2011

Oat (Avena sativa L.) quality improvement for increased beta-glucan concentration

Mark Newell
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Agronomy and Crop Sciences Commons

Recommended Citation
Newell, Mark, "Oat (Avena sativa L.) quality improvement for increased beta-glucan concentration" (2011). Graduate Theses and Dissertations, 10339.
http://lib.dr.iastate.edu/etd/10339

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Oat (*Avena sativa* L.) quality improvement for increased beta-glucan concentration

by

Mark Andrew Newell

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Plant Breeding

Program of Study Committee:
William D. Beavis, Co-major Professor
M. Paul Scott, Co-major Professor
   Dianne Cook
   Jode W. Edwards
   Pamela J. White

Iowa State University
Ames, Iowa
2011
Copyright © Mark Andrew Newell, 2011. All rights reserved.
DEDICATION

I would like to dedicate this dissertation to my Grandmother Twila Mae Blecha.
# TABLE OF CONTENTS

LIST OF TABLES ................................................................. vii

LIST OF FIGURES ................................................................. ix

ACKNOWLEDGEMENTS ............................................................... xiii

CHAPTER 1. GENERAL INTRODUCTION ........................................... 1
  1.1 Objectives ........................................................................... 1
  1.2 Rationale and Significance ................................................... 1
  1.3 Literature Review .............................................................. 1
    1.3.1 Oat ........................................................................... 1
    1.3.2 Evaluation of Oat $\beta$-Glucan ........................................ 2
    1.3.3 QTL Detection for Oat $\beta$-glucan .................................... 3
    1.3.4 Clustering High-Dimensional Data ..................................... 5
  1.4 Authors’ Roles ..................................................................... 5
  1.5 Dissertation Organization .................................................. 7

CHAPTER 2. POPULATION STRUCTURE AND LINKAGE DISEQUI-
            LIBRIUM IN OAT (AVENA SATIVA L.): IMPLICATIONS FOR GENOME-
            WIDE ASSOCIATION STUDIES .............................................. 8

  Abstract ................................................................................. 8
  2.1 Introduction ........................................................................ 9
  2.2 Materials and Methods ...................................................... 11
    2.2.1 Plant Material ............................................................ 11
    2.2.2 DArT Genotyping ....................................................... 11
    2.2.3 Data Curation ............................................................ 12
2.2.4 Model-Based Cluster Analysis ........................................... 13
2.2.5 Accounting for Population Structure .................................. 13
2.2.6 Linkage Disequilibrium .................................................. 14
2.2.7 Map Distances ............................................................. 14
2.3 Results ............................................................................... 15
  2.3.1 Data Curation .............................................................. 15
  2.3.2 Population Structure ..................................................... 16
  2.3.3 Cluster Relationships .................................................... 17
  2.3.4 Linkage Disequilibrium .................................................. 18
  2.3.5 Correlation of $r$ Between Clusters ................................. 21
2.4 Discussion ........................................................................... 23
  2.4.1 Population Structure ..................................................... 23
  2.4.2 Linkage Disequilibrium .................................................. 24
2.5 Acknowledgements ............................................................ 26

CHAPTER 3. MICRO-ENZYMATIC EVALUATION OF OAT (AVENA SATIVA L.) BETA-GLUCAN FOR HIGH-THROUGHPUT PHENOTYPING ................................................................. 28

Abstract ....................................................................................... 28
3.1 Introduction ............................................................................ 29
3.2 Materials and Methods ......................................................... 30
  3.2.1 Micro-Enzymatic $\beta$-glucan Assay ............................... 30
  3.2.2 Streamlined $\beta$-glucan Assay ...................................... 31
  3.2.3 Genetic Material ............................................................. 31
  3.2.4 Statistical Analysis .......................................................... 32
3.3 Results ................................................................................... 33
3.4 Discussion .............................................................................. 35
3.5 Acknowledgements ............................................................... 37
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>5.3.1 Simulated</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>5.3.2 Empirical</td>
<td>63</td>
</tr>
<tr>
<td>5.4</td>
<td>Results</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>5.4.1 Simulated Data</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>5.4.2 Empirical Data</td>
<td>71</td>
</tr>
<tr>
<td>5.5</td>
<td>Conclusion</td>
<td>75</td>
</tr>
<tr>
<td>5.6</td>
<td>Acknowledgements</td>
<td>76</td>
</tr>
<tr>
<td>6</td>
<td>Abstract</td>
<td>77</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>78</td>
</tr>
<tr>
<td>6.2</td>
<td>Materials and Methods</td>
<td>79</td>
</tr>
<tr>
<td>6.3</td>
<td>Techniques to Engage</td>
<td>82</td>
</tr>
<tr>
<td>6.4</td>
<td>Middle School Students</td>
<td>86</td>
</tr>
<tr>
<td>6.5</td>
<td>Closing Remarks</td>
<td>86</td>
</tr>
<tr>
<td>6.6</td>
<td>Acknowledgements</td>
<td>86</td>
</tr>
<tr>
<td>7</td>
<td>Conclusion</td>
<td>87</td>
</tr>
<tr>
<td>8</td>
<td>Bibliography</td>
<td>89</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

| Table 2.1 | Descriptions of the six oat clusters designated C1 to C6 identified using model-based cluster analysis. | 17 |
| Table 2.2 | Average pair-wise genetic distances between the six germplasm clusters. | 18 |
| Table 2.3 | Distribution of $r^2$ for the six germplasm clusters and the entire sample that included 1,205 lines. | 19 |
| Table 2.4 | Average correlation of LD (measured as $r$) between clusters at map distances of 0 to 5 (above diagonal) and $>5$ to 40 cM (below diagonal). | 23 |
| Table 3.1 | Summary of results for five methods of comparison including $P1:S1$, $P2:S2$, $LS:S1$, $P1:P2$, and $S1:S2$. $P1$ and $P2$ refer to the first and second replicates of the micro-enzymatic method, $LS$ refers to the least squared means for $P1$ and $P2$, and $S1$ and $S2$ refer to the first and second rep of the streamlined method. Results for sample size, Pearson and Spearman correlations, repeatability, and error variance are included for each comparison. | 35 |
| Table 4.1 | Results for the significant markers for the $P1K$, $P2K$ and $K$ models including the score, $R^2$, $R^2_{LR}$ and the FDR $q$-value obtained using the Benjamini and Hochberg method for multiple testing. The FDR cutoff for significance was 0.25. Marker names separated by a forward slash represent merged markers as a result of the data cleaning steps. | 52 |
Table 5.1 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. 66

Table 5.2 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. 68

Table 5.3 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. 72
LIST OF FIGURES

Figure 2.1 Scatter plots of principal component 1 (PC1) versus PC2 (a) and PC1 versus PC3 (b) showing the 1,205 lines making up the six clusters. Percent of the total variation accounted for by each PC is denoted in the axes titles in parentheses. .................. 16

Figure 2.2 Cluster dendrogram showing the genetic distances (measured as Manhattan distance) between clusters using wards linkage ............ 18

Figure 2.3 Scatter plot of LD ($r^2$) decay for all 2,225 linked marker pairs as a function of map distances (cM) for the 1,205 oat lines. The LD for 14,122 unlinked marker pairs is shown at an arbitrary distance of 42 cM. The inset graph shows the LD decay from 0 to 5 cM. ............. 20

Figure 2.4 Correlation of $r$ as a function of genetic distance between germplasm clusters at map distances of 0, >0 to 5, and >5 to 10 cM. Regression lines refer to the three map distances from top to bottom, respectively. 21

Figure 2.5 Average linkage disequilibrium (measured as $r^2$) at binned map distances for the three distance measures used in this study, PairD, AveD, and MapD. ................................. 22

Figure 3.1 $\beta$-glucan values enzymatically determined using the streamlined versus the micro-enzymatic method for P1:S1, P2:S1, and LS:S1 with 95% confidence bands. Pearson correlations for the three comparisons were significant with values of 0.85, 0.84, and 0.86, respectively. Spearman correlations for all three approaches were also significant with values of 0.81, 0.83, and 0.83, respectively. ............. 34
Figure 3.2  Results for the method of differences to determine the types of bias in the micro-enzymatic method compared to the streamlined approach with 95% confidence bands. \( P_1:S_1, P_2:S_1, \) and \( LS:S_1 \) all show fixed and proportional bias represented by the mean of values significantly greater than zero and the slope of the regression significantly greater than zero, respectively.  

Figure 4.1  Bar graph showing the countries of origin for the lines included in the study colored by the GRIN \( \beta \)-glucan classification. Counts are represented by the \( \log_{10}(\text{count})+1 \).  

Figure 4.2  Distribution of \( \beta \)-glucan BLUPs for the lines showing the bimodal distribution as a result of the selection process. Triangles beneath the distribution represent the \( \beta \)-glucan BLUPs for field checks including Buff, Excel, Winona, Cherokee, IA02130-2-2, Baker, HiFi, ND030288, and N979-5-1-22 from left to right, respectively.  

Figure 4.3  Scatterplot showing the relationship between the GRIN standardized \( \beta \)-glucan values and the \( \beta \)-glucan BLUPs based on two years with two replicates per year. The dashed line represents the regression between the two measures (correlation=0.68).  

Figure 4.4  Distribution of \( p \)-values for the six models included for evaluation of \( P \) in the mixed model. Axes represent the cumulative \( p \) versus the observed \( p \) in the \(-\log_{10}\) scale where the dashed line represents the Null. Models that do not include \( K \) (\( P_1, P_2, \) and \text{Simple} \) do not adequately account for false positives in contrast to \( P_1K, P_2K, \) and \( K \) that effectively reduce the number of false positives.  

Figure 4.5  Manhattan plots for the eight models used for association analysis showing the scores for each marker in clustered order calculated as \(-\log_{10}(p)\). Significant scores using an FDR of 0.25 are represented by bold points.
Figure 4.6  Arrangement of β-glucan rice candidates and oat DArT sequences on the rice genome. Rows represent the rice chromosome number and a plus symbol represents precise positions of the sequences. The entire length of each chromosome is not shown; high and low regions with no DArT sequence hits are not included.  

Figure 5.1  (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.  

Figure 5.2  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.  

Figure 5.3  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.  

Figure 5.4  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.
Figure 5.5  Diagnostic plot of Hubert’s gamma for varying number of clusters solutions to distinguish between the case when either one or two clusters are present, illustrated on simulated data. One, two and six cluster results are show, 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. 

Figure 5.6  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 $p < 0.01$. 

Figure 5.7  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.
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting this research. First and foremost Jean-Luc Jannink for his excellent guidance and expertise in research and his willingness to be a great mentor. My major advisors Bill Beavis and Paul Scott for providing me with the opportunity to pursue a PhD in plant breeding and the resources needed to conduct the research. Di Cook for her excellent teaching and collaboration that generated my interests in data exploration, visualization, and analysis. My committee members Jode Edwards and Pam White for their willingness to offer guidance in their respective disciplines. George Patrick for his support in the field. Symbi, Iowa’s NSF-GK-12, and the State Science and Technology Fair of Iowa for the opportunity to mentor the next generation of young scientists in these great programs. And lastly, my colleague Franco Asoro for his excellent friendship and support throughout this research.
CHAPTER 1. GENERAL INTRODUCTION

1.1 Objectives

The objectives of this research are to (1) implement and explore prospects for genome-wide association studies in oat for increased $\beta$-glucan content, (2) develop and test an alternative approach for clustering multivariate data, and (3) document techniques to engage young students in plant breeding and the sciences during outreach activities.

1.2 Rationale and Significance

Health issues related to diet have become a major problem facing developed countries over the past few decades. In many countries this problem is more emphasized in teens where levels of obesity and cases of type-2 diabetes have been on the rise. Consumption of oats as a whole grain has consistently been shown to have a positive impact on cholesterol, diabetes, and obesity. The positive health benefits of consuming oats as a whole grain food are attributed in part to $\beta$-glucan, a soluble fiber found in the cell walls. This dissertation explores and implements genetic approaches to increase $\beta$-glucan content in oat to increase its nutritional value.

1.3 Literature Review

1.3.1 Oat

The family POACEAE, commonly referred to as the grasses, is the largest plant family consisting of over 10,000 species worldwide (Kellogg, 2001). Cereals, grass species used for their
edible seeds, consist of six major crops including barley (*Hordeum vulgare* L.), wheat (*Triticum aestivum* L.), maize (*Zea mays* L.), rice (*Oryza sativa* L.), sorghum (*Sorghum bicolor* L.), and oat (*Avena sativa* L.). Maize production was highest worldwide in 2009 with 818 million metric tonnes (MT) produced, followed by wheat (686 million MT), rice (685 million MT), barley (152 million MT), sorghum (56 million MT), and oat (23 million MT). Oat production has been steadily decreasing in the US since 1955, mainly due to the profit decline compared to other crops such as corn and soybean (Murphy and Hoffman, 1992). Although there has been a steady decrease in oat production, the oat crop encourages crop diversity and control of plant diseases, insects, and weeds, and reduces soil erosion when included in rotations. It is therefore a useful crop from the perspective of agricultural sustainability. In addition, oat is highly recognized for its positive health benefits when consumed as a whole grain food. The positive health benefits associated with consumption of oats are attributed in part to mixed-linkage (1-3, 1-4)-β-D-glucan (referred to as β-glucan). β-glucan is a hemicellulose that makes up approximately 75% of the endosperm cell walls (Miller et al., 1995). Research on the role of oat β-glucan in the human diet has been consistent in showing health improvements with respect to blood pressure (Keenan et al., 2002), diabetes (Jenkins et al., 2002), cholesterol (Braaten et al., 1994), and the immune response (Estrada et al., 1997).

### 1.3.2 Evaluation of Oat β-Glucan

Throughout this dissertation β-glucan content and concentration are used synonymously. Two preferred methods for the determination of β-glucan content exist including the enzymatic and fluorimetric methods. The first description of the enzymatic method was by McCleary and Glennie-Holmes (1985) and was later modified to be faster and more efficient (McCleary and Codd, 1991). Currently, the modified method is sold as a streamlined mixed-linkage β-glucan kit (Megazyme Int., Wicklow, Ireland). Comparisons between the original method and the streamlined method yield a significant correlation coefficient of 0.98, indicating that the streamlined method is faster without sacrificing accuracy or precision. The procedure is as follows: 1) ground samples are suspended and hydrated in a buffer solution, 2) samples are incubated with lichenase enzyme and filtered, 3) the filtrate is then hydrolysed to completion.
with $\beta$-glucosidase, and 4) the D-glucose produced is assayed using a glucose oxidase/peroxidase reagent. The fluorometric method is based on the staining of $\beta$-glucan by a fluorescence enhancer Calcofluor (Vis and Lorenz, 1997). Although both methods perform well, the enzymatic method is the preferred method for grain flour $\beta$-glucan evaluation. Chernyshova et al. (2007) made additional modifications to the enzymatic method, which allowed an increase in the number of samples analyzed per unit time and a decrease in cost per sample. The modification made by Chernyshova et al. (2007) was use of a 96-well plate before adding the oxidase/peroxidase reagent in step four (outlined above) of the procedure.

1.3.3 QTL Detection for Oat $\beta$-glucan

As a result of the positive health implications associated with oat consumption, substantial breeding efforts have focused on increased $\beta$-glucan content. For the most part, selection has been based on phenotype where selection occurs within families produced by crossing two inbred lines (Cervantes-Martinez et al., 2001; Chernyshova et al., 2007). Although genetic gain has been substantial with phenotypic selection, marker development in oat enables selection that utilizes genotypic information for individuals. With the ultimate goal to improve the response to selection, the identification of individual loci that affect complex traits like $\beta$-glucan, referred to as quantitative trait loci (QTL), has been a major focus. Numerous linkage mapping studies for various traits have been performed that identify QTL in bi-parental populations in oat (e.g. Rines et al. (2006); Holland (2007)). The crossing of two parents generates linkage disequilibrium (LD; the nonrandom association of alleles at two loci) between linked loci in such populations. This experimentally-generated LD spans large chromosomal regions. Thus, any QTL for which the two parents differ will be in strong LD with observable DNA markers and if the QTL has a large effect the power to detect it will be high. Unfortunately, the results from such studies sample only two alleles per locus and result in QTL with low resolution (Kearsey and Farquhar, 1998).

An alternative approach for QTL detection are genome-wide association studies (GWAS). This approach is also dependent on the level of LD but, unlike linkage mapping, it depends on historical LD generated prior to any experimental work and broken down by many generations...
of recombination. Thus, chromosomal regions that remain in LD span shorter distances and have the potential to deliver high-resolution mapping. Additionally, because GWAS utilizes populations of lines unrelated by any specific crossing design (Zhu et al., 2008), the allelic diversity sampled is likely to be greater than two alleles per locus as is the case for linkage mapping. Due to the fact that the populations of lines used for GWAS are not developed experimentally, differential degrees of relationship, also called population structure, are likely present. For GWAS, population structure can result in false positives (Kennedy et al., 1992). A classic example of the effect of population structure on GWAS was conducted in humans where there was a strong negative association between a particular haplotype and type-2 diabetes in two Native American Tribes (Knowler et al., 1988). Although the haplotype was a candidate for disease, it was found that the haplotype was not a disease locus but a marker for Caucasian admixture. The conclusion was thus that the presence of the Caucasian alleles and the associated decrease of Native American alleles lowered the risk of disease, rather than that there was direct action of the haplotype itself.

A common statistical analysis for GWAS in plants is the mixed-model approach first described by Yu et al. (2006). The mixed-model approach accounts for both population structure fixed effects and pair-wise relationships, referred to as kinship, considered a random effect. Implementation of the mixed-model has been most useful for its ability to control for false positives associated with differential relationships. It has been used for various GWAS in crop plants including rice (Huang et al., 2010), wheat (Zheng et al., 2009), barley (Beattie et al., 2010), and maize (Beló et al., 2008). Since its initial publication, the mixed-model has been thoroughly tested with various modifications with respect to population structure and kinship (Stich et al., 2008).

The resolution for GWAS is not only dependent on the extent of LD but also the marker density. Recent advances in marker technology for oat have led to the development of Diversity Array Technology (DArT) markers (Tinker et al., 2009). Since DArT development, they have been utilized in genetic mapping of oat crown rust resistance gene Pc91 (McCartney et al., 2011), and genomic selection for oat β-glucan, heading date, groat percentage, plant height, and yield (Asoro et al., 2011). The marker density that has been gained with development
of the DArT marker system also enables the application of GWAS to detect QTL controlling β-glucan content in oat. Identification of QTL controlling β-glucan content has been limited in oat (Kianian et al., 2000; De Koeyer et al., 2004). Fortunately, a comparative genomics approach can give some insight into candidate genes from other grass species (Fincher, 2009). First identified in barley (Han et al., 1995) and later identified on chromosome seven of rice (Burton et al., 2006), the CslF gene family is known to have β-glucan synthase function. Thus, this gene family is an important candidate for QTL detection for increased β-glucan content in oat.

1.3.4 Clustering High-Dimensional Data

As molecular tools rapidly develop, data sets will become larger over time. Depending on the specific research question, clustering the data into meaningful groups, or subpopulations, maybe useful to uncover underlying structure. A first step in implementing this approach is to identify the number of clusters that exist. Although current methods, such as k-means and hierarchical clustering perform well they fail to address this fundamental question (Fraley and Raftery, 2003). On the other hand, model-based cluster analysis attempts to overcome this drawback by maximizing some information criterion. Currently the most common model-based method for clustering genetic data is implemented in the computer software STRUCTURE (Pritchard et al., 2000). In addition to the computational issues associated with STRUCTURE software (Vähä et al., 2007; Hamblin et al., 2010), it makes genetic assumptions that are rarely met in breeding populations. Although STRUCTURE does work well with respect to deciding the number of clusters, more common methods of cluster analysis (namely hierarchical) have also been shown to perform well for molecular marker data (Odong et al., 2011).

1.4 Authors’ Roles

The authors of Chapter 2 entitled “Population Structure and Linkage Disequilibrium in Oat (Avena sativa L.): Implications for Genome-Wide Association Studies” were Mark A. Newell, Dianne Cook, Nicholas A. Tinker, and Jean-Luc Jannink. Newell was the primary researcher and conducted all analyses, Cook played an integral role in data exploration/visualization
of the results and cluster analysis, Tinker was a corresponding author responsible for all of the distances and contributed substantially to the materials and methods and discussion, and Jannink was a corresponding author who set the research objectives and played a major role in conducting the research and writing of the manuscript.

The authors for Chapter 3 entitled “Micro-Enzymatic Evaluation of Oat (Avena sativa L.) $\beta$-glucan for High-Throughput Phenotyping” were Mark A. Newell, Hyun Jung Kim, Franco G Asoro, Adrienne Moran Lauter, Pamela J. White, M. Paul Scott, and Jean-Luc Jannink. Newell was the primary researcher responsible for all statistical analyses and micro-enzymatic evaluations, Kim was responsible for all of the streamlined evaluations, Asoro provided all of the genetic material used in the manuscript, Moran Lauter played an integral role in the laboratory evaluation for the micro-enzymatic approach, White contributed to the work conducted by Kim for the streamlined approach, Scott was the corresponding author and contributed in the development of the micro-enzymatic approach and writing of the manuscript, and Jannink contributed substantial effort in design and analyses of the manuscript.

The authors for Chapter 4 entitled “Genome-Wide Association Study for Oat (Avena sativa L.) Beta-Glucan using Germplasm of Worldwide Origin” were Mark A. Newell, Franco G. Asoro, M. Paul Scott, Pamela J. White, William D. Beavis, and Jean-Luc Jannink. Newell was the primary researcher and author responsible for all phenotypic and statistical analyses, Asoro contributed to field and laboratory experiments, Scott contributed substantial effort in evaluation of $\beta$-glucan, White contributed consultation for evaluation of $\beta$-glucan, Beavis consulted on the statistical analyses conducted, and Jannink set the research objectives and was the corresponding author for the manuscript.

The authors for Chapter 5 entitled “Determining the Number of Clusters for High-Dimensional Data with Application to Exploring Crop Population Structure” were Mark A. Newell, Dianne Cook, Heike Hofmann, and Jean-Luc Jannink. Newell was the primary and corresponding author and invented the algorithm, Cook contributed substantially to the entire manuscript, Hofmann contributed effort in coding of the analysis, and Jannink consulted on the approach.

The author for Chapter 6 entitled “Outreach Education - Techniques to Engage Middle School Students” was Mark A Newell. Newell was the primary and corresponding author of
1.5 Dissertation Organization

The dissertation has seven chapters including this general introduction (Chapter 1), five body chapters (Chapters 2-6), and a general conclusion (Chapter 7). Each of the five body chapters is a separate manuscript that is either published, submitted for publication, or soon to be submitted excluding the general introduction (Chapter 1) and conclusion (Chapter 7). Chapter 2 is a survey of the population structure and linkage disequilibrium in oat and its implications for genome-wide association studies. Chapter 3 is a proposed method for high-throughput phenotyping for \( \beta \)-glucan evaluation. Chapter 4 is a genome-wide association study to identify quantitative trait loci associated with increased \( \beta \)-glucan content in a world collection of oat germplasm. Chapters 2, 3, and 4 address Objective 1 of the dissertation. Chapter 5 presents an algorithm to identify the number of clusters in multivariate data and addresses Objective 2 of the dissertation. Chapter 6 is a case study that presents techniques to engage middle school students during outreach activities and addresses Objective 3 of the dissertation. Chapter 7 is the general conclusion giving a brief outline of the research as a whole.
CHAPTER 2. POPULATION STRUCTURE AND LINKAGE DISEQUILIBRIUM IN OAT (AVENA SATIVA L.): IMPLICATIONS FOR GENOME-WIDE ASSOCIATION STUDIES

A paper published in Theoretical and Applied Genetics\(^1\)

Mark A Newell\(^2\), Dianne Cook\(^3\), Nicholas A Tinker\(^4\), and Jean-Luc Jannink\(^5\)

Abstract

The level of population structure and the extent of linkage disequilibrium (LD) can have large impacts on the power, resolution, and design of genome-wide association studies (GWAS) in plants. Until recently, the topics of LD and population structure have not been explored in oat due to the lack of a high-throughput, high-density marker system. The objectives of this research were to survey the level of population structure and the extent of LD in oat germplasm and determine their implications for GWAS. In total, 1,205 lines and 402 diversity array technology (DArT) markers were used to explore population structure. Principal component analysis and model-based cluster analysis of these data indicated that, for the lines used in this study, relatively weak population structure exists. To explore LD decay, map distances of 2,225 linked DArT marker pairs were compared with LD (estimated as \(r^2\)). Results showed that LD between linked markers decayed rapidly to \(r^2 = 0.2\) for marker pairs with a map distance of 1 centimorgan (cM). These findings have important implications for GWAS in oat, suggesting that LD decay is rapid and that marker spacing should be considered when designing association studies.

---

\(^1\)Reprinted with permission of Theoretical and Applied Genetics, 2011, 122, 623-632
\(^2\)Primary researcher and author, Department of Agronomy, Iowa State University, Ames, IA 50011, USA
\(^3\)Department of Statistics, Iowa State University, Ames, IA 50011, USA
\(^4\)Author for Correspondence, Agriculture and Agri-Food Canada, ECORC, 960 Carling Ave, Ottawa, ON K1A OC6, Canada
\(^5\)Author for Correspondence, USDA-ARS, Robert W. Holley Center for Agriculture and Health, Cornell University Department of Plant Breeding and Genetics, 407 Bradfield Hall, Ithaca, NY 14853, USA
distance of 1.0 centi-Morgan (cM). For GWAS, we suggest a minimum of one marker every cM, but higher densities of markers should increase marker-QTL association and therefore detection power. Additionally, it was found that LD was relatively consistent across the majority of germplasm clusters. These findings suggest that GWAS in oat can include germplasm with diverse origins and backgrounds. The results from this research demonstrate the feasibility of GWAS and related analyses in oat.

2.1 Introduction

Oat (Avena sativa L.) is a grass species grown as a grain or forage crop predominantly in temperate short-season regions. Oat has received significant attention in recent years due to the human health benefits of consuming it as a whole-grain food. There are many active oat breeding programs around the world where improved oat varieties are developed through phenotypic selection for complex traits such as disease resistance, yield, lodging, and stress tolerance. For example, there are at least 12 publicly funded oat breeding programs in the USA and Canada. Recently, methods to improve the response to selection have focused on the identification of individual loci, known as quantitative trait loci (QTL), controlling these complex traits. There are numerous QTL mapping studies that have utilized linkage-based analysis of bi-parental populations in oat (e.g. reviewed by Rines et al. (2006); Holland (2007)). Although this has proven to be a powerful approach for QTL detection, it delivers low-resolution, population specific QTL, and samples only a small portion of the allelic diversity present in the germplasm available (Zhu et al., 2008). Genome-wide association studies (GWAS) attempt to overcome the pitfalls associated with linkage mapping in bi-parental populations. Genome-wide association studies have the ability to identify useful allelic diversity and to map this diversity with high resolution within complex plant pedigrees that are typical of breeding programs (Jannink et al., 2001). From a practical perspective, GWAS have been applied in many grain crops, including rice, maize, barley, and wheat (Agrama et al., 2007; Beló et al., 2008; Kraakman et al., 2006; Zheng et al., 2009). Implementation of GWAS in oat for QTL detection could be valuable to the oat community.

The ability of GWAS to deliver high-power, high-resolution results is largely dependent
on the extent of linkage disequilibrium (LD) within the working population. Also known as gametic phase disequilibrium, LD is defined as the non-random association of alleles at two loci (Falconer and Mackay, 1996) and is affected by mutation, admixture, selection, drift, population structure associated with breeding history, and reproductive biology (reviewed by Flint-Garcia et al. (2003)). Additionally, since the mechanisms mentioned may differentially affect different genomic regions, this can introduce LD heterogeneity across the genome. This makes the power and resolution achieved in GWAS highly dependent on the species and the population being evaluated.

The extent of LD among the common grass species varies with respect to the crop and the population chosen for evaluation. For example, in maize, an allogamous species, LD decays over relatively shorter distances compared with autogamous crops. Remington et al. (2001) reported that LD (measured as $r^2$) declined to 0.1 within 1,500 basepairs (bp) in a set of 102 maize inbred lines representing breeding germplasm from temperate and tropical regions. Tenaillon et al. (2001) found similar results in a group of 25 maize lines consisting of 16 landraces and nine elite inbreds: LD decayed to 0.15 at 500 bp for the combined dataset. Unlike maize, barley is a self-pollinated crop with strong population structure due to variation in growth habit and kernel row number. Zhang et al. (2009) reported that LD extends to 2.6 cM at $r^2$ equal to 0.2 for a group of elite Canadian lines. This is in agreement with Hamblin et al. (2010) in a study of North American elite germplasm. Similar to barley, oat is a self-pollinated species; thus it is expected that LD decay will occur over relatively long map distances.

In addition to LD, the power and resolution of GWAS is also dependent on marker density. Until recently, the lack of genetic markers and a method to deliver high-throughput genotyping have limited the options for identifying QTL in oat. Diversity Array Technology (DArT) markers developed recently for oat have greatly increased the density of available markers (Tinker et al., 2009). These markers were developed based on random clones isolated from 60 elite varieties of diverse global origin, making them useful in diversity analysis as well as in linkage mapping. Since they can be applied in parallel using a cost effective assay, they show good potential for use in QTL detection, comparative mapping, marker-assisted selection (MAS), and genomic selection.
Due to the fact that the power and resolution of GWAS depends greatly on the extent of LD across the genome, it is important to survey this extent. In this study, we determine and discuss the population structure and the extent of LD among DArT markers in an extensive worldwide collection of oat consisting of varieties, breeding lines, and landraces.

2.2 Materials and Methods

2.2.1 Plant Material

Datasets from four independently assembled germplasm collections were combined in this study to increase the diversity and representation of the results. The four component datasets consisted of 462, 466, 198, and 279 lines, each set having been assembled for a variety of other purposes that will be published elsewhere. The set of 462 lines consisted of current North American elite varieties. The set of 466 lines was a world collection of oat germplasm from the Germplasm Resources Information Network (USDA-ARS, National Genetic Resources Program, 2010) consisting of varieties, breeding lines, and landraces. The set of 198 lines consisted of varieties of global origin that were used by Tinker et al. (2009) in the initial DArT development work, and the set of 279 lines was an extension of this set intended for use in association mapping. The combined dataset represented a total of 1,405 lines from 53 countries.

2.2.2 DArT Genotyping

Plants were grown under greenhouse conditions and tissue was collected from single plants or from multiple bulked seedlings originating from seed of a single oat panicle. Isolation of DNA was performed by a variety of methods in use by different collaborating laboratories. It was already evident in the work of Tinker et al. (2009) that these different extraction methods would not affect the DArT assays. DArT marker analysis was performed by Diversity Arrays P/L, Canberra, Australia using methods described by Tinker et al. (2009). Due to the fact that DArT markers were under development during the initial stages of this work, the four datasets submitted for DArT genotyping provided datasets with varying numbers of markers.
2.2.3 Data Curation

Because four independently assembled component datasets were merged, some duplicate lines existed. Previous research has suggested that this could contribute to biased estimates of LD (Breseghello and Sorrells, 2006). In addition to duplicate lines, the datasets ranged in numbers of markers from 1,001 to 1,958; thus a core marker set was required. Data curation was implemented to accomplish the following: (1) identify and merge lines that were submitted more than once across the datasets, (2) identify and merge redundant markers, and (3) eliminate lines and markers with insufficient data points.

The DArT marker assay produces dominant marker scores, which were represented as a matrix of 1s and 0s. Genetic distances (measured as Manhattan distance) were calculated across all genotype pairs and expressed as the proportion of the maximum. Pairs of lines with genetic distances <5% that had similar names were merged, retaining the line with the most complete data. Pairs with genetic distances <2% that shared at least 200 markers in common were also merged regardless of nomenclature. The rationale for this was that even if the two lines are in fact distinct, they must be strongly related, and that representing them as a single entry would be more meaningful in the determination of LD.

After removal of redundant lines, redundant markers were also removed using a similar fashion. Markers based on DArT clones having DNA sequence data represented in the same contiguous DNA assembly (Tinker et al., 2009), and with scores that differed by <2% were merged. When two markers belonged to the same sequence assembly, but differed by more than 2%, the marker with most missing data was removed. This approach assumes that all markers belonging to the same sequence assembly are identical. Markers with identical scores were also merged if the scores were non-ambiguous across 100 or more lines. This merging process was performed for compatibility with the mapping data set and the resulting linkage maps. This process resulted in the merging of only four pairs of markers (1% of the total marker number) and is unlikely to have affected the results because such markers would be at distance zero and in perfect disequilibrium. Although these assumptions of marker identity may occasionally be incorrect, these procedures were selected as a conservative approach to remove redundant
markers that would otherwise cause an upward bias in the estimate of LD for the short artificial linkage intervals caused by slight variations in the scoring of identical markers.

Because many of the markers were not scored on all of the datasets, markers that were scored in fewer than 80% of the lines were removed, followed by removal of all monomorphic markers. These final steps ensured that only markers and lines with a sufficient amount of data were retained. Genetic distances among all lines were re-computed after these final data curation steps. The resulting dataset after data curation consisted of 1,205 lines and 402 markers.

2.2.4 Model-Based Cluster Analysis

The \texttt{Mclust} package (Fraley and Raftery, 2006) in the statistical software R was used to identify clusters among lines. A model-based approach was used because it determines the number of clusters and cluster membership simultaneously and it does not have underlying genetic assumptions that are rarely met. The package identifies the optimal model according to the Bayesian Information Criterion (BIC) for expectation-maximization (EM) initialized by hierarchical clustering for parameterized Gaussian mixture models (Fraley and Raftery, 2007). Due to the large number of dimensions (402 markers), cluster analysis was implemented on the principal components. By using the principal components instead of the marker data, it was possible to fit models of varying shape, size, and orientation. Models with between 2 and 30 clusters were compared.

2.2.5 Accounting for Population Structure

Principal components analysis (PCA) was implemented to account for population structure (Price et al., 2007). First, missing marker values were replaced by the mean for the marker. PCA was applied to the lines using the \texttt{prcomp} function in the statistical software R, which adequately handles computational issues of high-dimensional data. The choice of the number of principal components used was based on the scree plot of eigenvalues (Cattell, 1966). Singular value decomposition was used to account for population structure using the appropriate number of principal components from above. A matrix representing marker scores expected on the basis of population structure was calculated as $R = UDV'$, where $U$ is a matrix of left
singular vectors, $D$ is a diagonal matrix of singular values, and $V$ is a matrix of right singular vectors. The population structure matrix ($R$) was subtracted from the marker data, and LD was calculated as described below.

Population structure was accounted for in all LD calculations using the aforementioned approach. For the six clusters (percent of the variation shown in parentheses), four (23.4), three (18.1), three (16.2), four (24.8), four (22.6), and seven (62.2) principal components were used, respectively. For the entire sample, five principal components (22.8) were used to account for population structure.

### 2.2.6 Linkage Disequilibrium

Three common methods for calculating LD exist in plants, denoted by $D$, $D'$, and $r^2$. For this research, the correlation squared ($r^2$) was used because (1) it is not as highly influenced by small sample sizes and low allele frequencies (Flint-Garcia et al., 2003), and (2) it is relevant for QTL mapping because it relates the amount of variance explained by the marker to the amount of variance generated by the associated QTL (Zhu et al., 2008). The calculation used is as follows: $r^2 = \left[ \frac{\sigma_{XY}}{\sigma_{X}\sigma_{Y}} \right]^2$, where $\sigma_{XY}$ is the covariance between marker $X$ and marker $Y$, and $\sigma_{X}$ and $\sigma_{Y}$ are the standard deviations for marker $X$ and marker $Y$, respectively. This calculation was applied using the R statistical package (R Development Core Team, 2009) to each marker pair using all available data points. The functional relationship between LD and map distance was determined by fitting the nonlinear model (Sved, 1971) $r^2 = \frac{1}{1 + 4ad}$, where $d$ is the map distance in cM and $a$ is an estimated regression coefficient. The parameter $a$ can also be interpreted as the effective population size of the population to which the analysis was applied.

### 2.2.7 Map Distances

The current lack of a consensus map in oat presents the issue of deciding on a map distance measure that will adequately describe the LD decay. Most commonly, the LD for a pair of markers is compared to a map distance that is taken directly from a consensus map. However, the only map on which a large number of markers have been resolved is the updated Kanota
x Ogle map (Tinker et al., 2009), where there are approximately twice as many linkage groups as there are chromosomes in oat. In order to avoid bias and artifacts introduced by the map, it was decided to use direct counts of recombination events between each available pair of markers in the published Kanota x Ogle mapping data as the primary measure of map distance. These recombination estimates were expressed as centi-Morgan (cM) distances using the Kosambi mapping function and are identified hereafter as \( \text{PairD} \). For example, if two markers \( A \) and \( C \) showed a direct pair-wise map distance of 20 cM, this distance would be used in estimating LD decay regardless of what their distance was on the resolved map, even if they were not resolved to the same linkage group on the published map. Markers at distances of greater than 40 cM were considered unlinked. In order to compare this approach to the more common approach based on a resolved linkage map, we reconstructed the same analysis using resolved cM distances (identified as \( \text{MapD} \)) from the published map, which included 665 linked pair-wise LD measurements. Furthermore, since other unpublished map data were available, and since this would allow estimation of a greater number of pair-wise marker distances, we tested a third additional approach: estimates of recombination between each pair of markers were computed from all available mapping populations, and these were averaged (represented hereafter as \( \text{AveD} \)). Pairs of markers were excluded under the following situations: (1) when the minimum distance was less than or equal to 5 cM but individual estimates varied by more than 10 cM and (2) when the minimum distance was greater than 5 cM and less than or equal to 20 cM and the distances varied by more than 200% of minimum. Pairs with \( \text{AveD} \) greater than 40 cM were considered unlinked.

### 2.3 Results

#### 2.3.1 Data Curation

The final dataset contained 1205 lines and 402 markers (additional data for the final dataset are given in Online Resource 1). Lines from the United States, Canada, and Germany were highly represented, accounting for 44, 16.5, and 4.6% of the lines in the study, respectively. Other countries represented by greater than 2% of the lines were Sweden, Turkey, the United
Kingdom, and the Russian Federation. The remaining 46 countries accounted for 19.3% of the lines included in the study. Some lines had multiple origins upon merging those with similar genotype that varied in origin, and these accounted for 2.2% of the lines. A further 2.8% of the lines had unknown origin. In total, there were 2,225 linked and 15,541 unlinked (>40 cM) pair-wise LD estimates used for the primary analysis of this study (PairD).

2.3.2 Population Structure

Principal component analysis and model-based cluster analysis were used to explore population structure. Cluster analysis was implemented on the first five principal components explaining 22.8% of the variation. The BIC for the different models were similar beyond six clusters but a defined peak was not present; therefore, the decision was made to use only six clusters (C1-C6). All of the clusters can be separated in the first three principal components that account for 8.8, 4.91, and 3.6% of the variation, respectively (Figure 2.1). Although the clusters can be separated by the first three PCs, the data represent a cloud in space where distinct clusters are not readily seen.

![Figure 2.1](image.png) Scatter plots of principal component 1 (PC1) versus PC2 (a) and PC1 versus PC3 (b) showing the 1,205 lines making up the six clusters. Percent of the total variation accounted for by each PC is denoted in the axes titles in parentheses.
The number of lines per cluster ranged from 20 to 334 (Table 2.1). Geographic origins of lines making up >5% of a cluster were evaluated to assess the relationship between the clusters and their origins. Because US and Canadian lines account for 60% of the 1,205 lines in the study, it was expected that these countries would make up a substantial amount of all clusters. In general, clusters C1 and C2 contain mainly Canadian lines, and clusters C3, C4, and C5 contain mainly US lines. Although it is difficult to distinguish the clusters by origin due to the high frequency of US and Canadian lines, clusters could be associated with important released oat lines. These include Assiniboia, Triple Crown, Jay, Kanota, and Ogle for clusters C1 to C5, respectively. Cluster C6 was a small cluster consisting of red oats (*A. sativa* ssp. *byzantina* K. Koch), typically grown as winter oats in the southern US.

Table 2.1 Descriptions of the six oat clusters designated C1 to C6 identified using model-based cluster analysis.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of lines</th>
<th>Origins$^a$</th>
<th>Representative line$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>166</td>
<td>CA, US</td>
<td>Assiniboia</td>
</tr>
<tr>
<td>C2</td>
<td>334</td>
<td>DE, US, SE, CA, RU, YU</td>
<td>Triple Crown</td>
</tr>
<tr>
<td>C3</td>
<td>209</td>
<td>US, CA, Unk</td>
<td>Jay</td>
</tr>
<tr>
<td>C4</td>
<td>184</td>
<td>US, TR, UK, AU</td>
<td>Kanota</td>
</tr>
<tr>
<td>C5</td>
<td>292</td>
<td>US, CA</td>
<td>Ogle</td>
</tr>
<tr>
<td>C6</td>
<td>20</td>
<td>US, CA, UK, Mul</td>
<td>Red oats</td>
</tr>
</tbody>
</table>

$^a$ Origins are identified as countries from which >5% of the lines in a cluster originated.

$^b$ An arbitrary but widely recognized line was selected to represent each cluster.

2.3.3 Cluster Relationships

Differences exist in the pair-wise relationships between clusters as seen in the PC scatter plots. Quantitative results for genetic distances (measured as Manhattan distance) between clusters are shown in Table 2.2; a graphical representation is shown in Figure 2.2. C6, a small group of winter red oats, is most distant from all other clusters, with an average distance of 138, whereas C3 is most closely related to all other clusters with an average distance of 62. All other clusters have similar average distances to other clusters in the range of 74-91. Interestingly, the genetic distances are in agreement with the origin and/or adaptation of the representative lines.
(Table 1) falling within each cluster. That is, C3 and C5 are most closely related (24) and are defined by Jay and Ogle, Indiana and Illinois lines, respectively, from the USA. The next most closely related are C1 and C2 (28), which contain Assiniboia and Triple Crown. Assiniboia was bred in Manitoba, Canada, while Triple Crown originated in Sweden but is released in, and highly adapted to, Western Canada. Last, C4 and C6 have a pair-wise genetic distance of 73; these clusters contain Kanota (C4), a winter red oat from Kansas and another small group of mostly red oat lines (C6). These results suggest that clustering was efficient in separating major lines and oat types for the germplasm used in this study.

Table 2.2 Average pair-wise genetic distances between the six germplasm clusters.

<table>
<thead>
<tr>
<th></th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>28</td>
<td>70</td>
<td>80</td>
<td>45</td>
<td>157</td>
</tr>
<tr>
<td>C2</td>
<td>-</td>
<td>43</td>
<td>82</td>
<td>75</td>
<td>165</td>
</tr>
<tr>
<td>C3</td>
<td>-</td>
<td>36</td>
<td>24</td>
<td>138</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>-</td>
<td>98</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>-</td>
<td>155</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.2 Cluster dendrogram showing the genetic distances (measured as Manhattan distance) between clusters using wards linkage

2.3.4 Linkage Disequilibrium

The extent of LD in a species determines the power and resolution of GWAS. For oat, it was expected that decay of LD would be over relatively long map distances because of its breeding history and reproductive biology. One way of summarizing breeding history is by estimating
the effective population size of the sample analyzed. For this study, the effective population size for the entire sample estimated from non-linear regression of $r^2$ on map distance was 92 with a standard error of six. In other words, an ideal randomly mating population of 92 diploid individuals (Falconer and Mackay, 1996) would be expected to have a similar rate of LD decay as the global oat population. Based on decay of LD in barley (Hamblin et al., 2010) and wheat (Chao et al., 2007), this effective population size appears typical of elite cultivated small grains.

For the 2,225 linked marker pairs used in this study, LD decays such that $r^2$ is equal to 0.1 at 2.5 cM (Figure 2.3). Within clusters, C1 to C5 show similar trends in LD decay, but quantitative differences can be seen (Table 2.3). In C6, the small cluster of mainly red oat lines, LD decays sporadically most likely due to the small sample size (Table 2.3). Cluster C5 had a relatively slower LD decay compared with the entire sample. A possible reason for this result could be the frequent use of the variety Ogle in this group as a parent in crosses for variety development (Figure 2.4).

Table 2.3 Distribution of $r^2$ for the six germplasm clusters and the entire sample that included 1,205 lines.

<table>
<thead>
<tr>
<th>Map distance (cM)</th>
<th>Linkage disequilibrium ($r^2$)</th>
<th>Entire sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cluster</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1</td>
<td>C2</td>
</tr>
<tr>
<td>0</td>
<td>0.280</td>
<td>0.221</td>
</tr>
<tr>
<td>&gt;0 to 5</td>
<td>0.164</td>
<td>0.093</td>
</tr>
<tr>
<td>&gt;5 to 10</td>
<td>0.087</td>
<td>0.043</td>
</tr>
<tr>
<td>&gt;10 to 40</td>
<td>0.015</td>
<td>0.009</td>
</tr>
<tr>
<td>Unlinked (&gt;40)</td>
<td>0.009</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Two alternate distance measures were used to determine if decay was dependent on the choice of distance. The first, MapD, was based on map distances in the Kanota x Ogle genetic map, and the second, AveD, was based on average recombination distances from six populations of oat, each composed of 75-150 recombinant inbred lines at generations 4-6 (unpublished data). MapD and AveD were composed of 665 and 3,228 linked pair-wise LD measurements, respectively. Both alternate measures of map distances showed similar trends compared with PairD, the primary distance used for analysis in this study, with some minor differences (Figure 2.5). MapD resulted in a slower decay that was more sporadic than both PairD and AveD.
Figure 2.3  Scatter plot of LD ($r^2$) decay for all 2,225 linked marker pairs as a function of map distances (cM) for the 1,205 oat lines. The LD for 14,122 unlinked marker pairs is shown at an arbitrary distance of 42 cM. The inset graph shows the LD decay from 0 to 5 cM.

However, AveD resulted in a curve very similar to PairD, suggesting that decay is not highly dependent on the distance measurement.

Identification of disequilibrium between unlinked marker pairs can be useful since these markers can affect GWAS. Unlinked marker pairs with high disequilibrium could indicate that unknown linkages or pseudo-linkages are present in oat. Quantitative results of LD between unlinked marker pairs are shown in Table 2.3 and are expectedly low. The average LD between unlinked markers is lower for the entire sample (0.004) than it is within each germplasm cluster (0.006-0.075). The estimate of LD between unlinked marker pairs is shown in Figure 2.3 at an arbitrary distance of 42 cM. The LD for these marker pairs are below an $r^2$ value of 0.1 except for one point that has a value of 0.15.
2.3.5 Correlation of $r$ Between Clusters

The design of GWAS depends on the consistency of gametic or LD phase across germplasm clusters. To address this, the correlation of LD (measured as the correlation of $r$) between cluster pairs was explored at map distances of zero, greater than zero to five, and greater than five to 10 cM (Figure 2.4). The correlation of $r$, rather than $r^2$, was used because $r$ has a signed value, making it more relevant to discuss consistency. Ten thousand permutations of the analysis of variance (ANOVA) were used to test for a significant relationship between correlation of $r$ and the genetic distance for varying map distances. There was a significantly negative relationship between the correlation of $r$ and the genetic distance for all map distance intervals ($P = 0.01$). These results match our expectations as well as previous results found in barley (Hamblin et al., 2010). However, when cluster C6 is removed from the analysis, there was no longer a significant relationship at any of the three map distance intervals. Quantitative
Figure 2.5  Average linkage disequilibrium (measured as $r^2$) at binned map distances for the three distance measures used in this study, PairD, AveD, and MapD.

Differences in consistency of LD between the six germplasm clusters from map distances 0 to 5 and >5 to 40 cM are shown in Table 2.4. At the interval of 0-5 cM, the average correlation of $r$ between clusters is 0.81, ranging from 0.71 to 0.92 when cluster C6 was excluded. Pairwise correlations of $r$ for cluster pairs that included C6 were lower, ranging from 0.32 to 0.49, which would be expected given the small sample size of this cluster ($n = 20$) and its more distant relationship with all other clusters (Table 2.2). For the interval of >5 to 40 cM, the average correlation of $r$ for cluster pairs was relatively lower with a value of 0.20. Given the weak population structure in oat that was described earlier, one would expect LD phase to be consistent across germplasm, as was shown here. These results indicate that LD is consistent across most oat germplasm, most notably at short-range map distances up to 5 cM where marker-QTL associations are most likely to occur.
Table 2.4  Average correlation of LD (measured as $r$) between clusters at map distances of 0 to 5 (above diagonal) and $>5$ to 40 cM (below diagonal).

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>-</td>
<td>0.74***</td>
<td>0.79***</td>
<td>0.80***</td>
<td>0.83***</td>
<td>0.32***</td>
</tr>
<tr>
<td>C2</td>
<td>0.24***</td>
<td>-</td>
<td>0.71***</td>
<td>0.77***</td>
<td>0.71***</td>
<td>0.49***</td>
</tr>
<tr>
<td>C3</td>
<td>0.24***</td>
<td>0.33***</td>
<td>-</td>
<td>0.89***</td>
<td>0.92***</td>
<td>0.37***</td>
</tr>
<tr>
<td>C4</td>
<td>0.21***</td>
<td>0.25***</td>
<td>0.30***</td>
<td>-</td>
<td>0.89***</td>
<td>0.39***</td>
</tr>
<tr>
<td>C5</td>
<td>0.39***</td>
<td>0.19***</td>
<td>0.39***</td>
<td>0.29***</td>
<td>-</td>
<td>0.36***</td>
</tr>
<tr>
<td>C6</td>
<td>0.09**</td>
<td>0.06*</td>
<td>-0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>-</td>
</tr>
</tbody>
</table>

*** Significant at 0.0001
** Significant at 0.01
* Significant at 0.05

2.4 Discussion

2.4.1 Population Structure

The level of population structure in a species has implications on the design and analysis of GWAS. For this study, PCA and cluster analysis were used to explore the amount of population structure. Major population structure in an organism can be observed by plotting the first few PCs (Menozzi et al., 1978; Price et al., 2007). For oat, examination of the first three principal components indicates that there is weak population structure within the germplasm evaluated. In contrast, barley is known to have strong population structure due to 2-row, 6-row, spring, and winter types (Hamblin et al., 2010). Plotting of the first two PCs for barley results in non-overlapping, distinct clusters similar to what would be expected for a species with strong population structure (Malysheva-Otto et al., 2006; Zhang et al., 2009). Oat also has four recognizable types, including naked, hulled, spring, and winter. Interestingly, although it was not tested because the majority of lines included in the study were spring, hulled types or unknown, the observed population structure is most likely independent of these types, other than for the small but distinct group of red winter varieties in cluster C6. Possible reasons for this could include different practices in exchange of germplasm among breeding programs, or breeding methods that more frequently utilize crosses among different oat types. Fortuitously, the relatively weak population structure of oat reported here suggests that GWAS can successfully span a wide diversity of oat types.
2.4.2 Linkage Disequilibrium

For this study, LD decay was explored for a set of germplasm consisting of 1,205 lines, and for six derived germplasm clusters identified using model-based cluster analysis. The amount of LD in the combined population decayed at a rate very similar to that within the derived germplasm clusters. Theory developed to predict LD in the presence of population structure suggests that overall LD should be similar to sub-population LD when extensive migration occurs between sub-populations (Sved, 2009). When migration is high, LD within the overall population should behave as in an unstructured, large population, as is the case for oat. While we have no records of the frequency of crosses made by breeding programs between the clusters that we found, we expect that they do occur, given the low level of differentiation between clusters. In contrast, barley is known to have strong population structure and shows large differences between germplasm clusters with respect to LD decay (Hamblin et al., 2010; Zhang et al., 2009). Most importantly, the design of GWAS depends on the consistency of gametic or LD phase across germplasm clusters. If the phase of LD differs among clusters, independent GWAS need to be conducted within each cluster. It is expected that LD for closely linked markers will be most similar between closely related clusters and that this similarity will decrease more slowly at greater distances in closely related clusters than it will in more distant clusters. Our results demonstrated that this relationship was significant only if cluster C6 was included. A possible reason for this result is that the remaining clusters are not genetically distinct enough for a large change in the correlation of \( r \) to occur. The consistency of LD phase for oat across most germplasm clusters identified here indicates that GWAS can include germplasm with diverse origins and backgrounds.

Using alternate estimates of map distance taken directly from the map (MapD), a similar but slightly slower and less consistent rate of LD decay was observed. The Kanota x Ogle genetic map is incomplete and contains more linkage groups than the 21 chromosomes in oat. Therefore, it is possible that some of the inconsistency in decay is due to the much smaller number of marker pairs for which an estimate of map distance was available. However, the slower rate of decay observed using MapD may also result from ubiquitous errors in map
construction. For example, if markers A and C are separated by several other markers on the map, and some mis-scoring has occurred in these intervening markers, then the map distance between markers A and C will be artificially stretched, resulting in an apparent slower decay in LD at longer map distances. For this reason, we think that the estimates based on PairD provide a more accurate representation of LD decay.

One concern for this study and for future research is the sub-optimal application of dominant bi-allelic markers in LD and GWAS studies, since a single dominant or recessive allele class can include genotypes with multiple alleles at a target locus. When this occurs, the ability to detect associations with a specific target allele is weakened. This problem will also apply to the use of bi-allelic SNP markers, except that the co-dominant nature of SNPs can eliminate confounding effects of residual heterozygosity. DArT and SNP markers may also differ in that a variety of mutation mechanisms can lead to a change of DArT allele (e.g., both point mutation in the DArT restriction site or a long insertion into the DArT fragment could lead to loss of fragment amplification), whereas generally only point mutations will cause a change of SNP allele. DArT marker mutation rates may therefore be higher than SNP mutation rates (which are known to be very low). The implications in the current study are that LD values are probably underestimated relative to what they would be if multi-allelic markers were available. The use of dominant or bi-allelic markers may become more powerful if multi-locus haplotypes can be used for GWAS, especially if higher-density SNP resources are developed at a later date.

Another source of error may arise due to the presence of markers that segregate as multiple loci but are identified by a single marker name. When this occurs, the dominant alleles of two or more loci will be confounded in the set of diversity data, even though a single locus may segregate normally within a given mapping population. This scenario is known to exist for DArT markers in other polyploid species such as wheat and likely exists to some degree in oat. Based on indirect estimates, the frequency of markers that segregate as multiple loci in oat is potentially about 5% (Tinker et al., 2009). Thus, two markers identified as being unlinked in Kanota x Ogle could be linked in a portion of the diversity panel, or vice versa. To address this, we have tested a third set of estimates for map distance that were derived from averaging the recombination fractions across multiple mapping populations. Although this analysis provided
a greater number of data points, the results were highly similar to the primary analysis using PairD from Kanota x Ogle, demonstrating that the estimates of LD decay are quite robust and probably not influenced by segregation of duplicate markers.

The power and resolution of GWAS is dependent on the extent of LD in a given population, assuming a suitable marker system is in place. In practice, $r^2$ between a marker and a QTL is equal to the percent of phenotypic variation of a QTL that can be explained by a marker. For oat, LD was on average 0.2 for DArT markers separated by 1.0 cM. Thus, the results from this study indicate that a marker every cM (2,000 markers total) would explain, on average, 20% of QTL variance. Since a marker and a QTL must also have similar minor allele frequencies to be in LD, we suggest that the number of markers should be on the order of 10,000 to increase the probability of identifying a marker that is in high LD with a QTL. At the current rate of development for oat DArT markers, this marker density is approachable in the near future. The authors are also engaged with collaborators in the development of new SNP marker resources for oat, a process that has been greatly assisted by the use of DArT markers to select diverse germplasm. However, these results do not imply that GWAS cannot succeed at much lower densities. The markers employed in this study are somewhat clustered, and since DArT markers are designed to target gene-rich regions, it is possible that many of these will be the same clusters that contain QTL. Many linkage blocks of favorable QTL alleles may also have been deliberately or inadvertently selected in breeding programs, and these same linkage blocks will contain markers in high LD. Therefore, while we encourage the development of high-density maps for GWAS, we do not discourage the exploration and utilization of QTL associations using existing molecular tools. Most importantly, this work demonstrates the distances at which LD can be expected, and the non-dependence of LD on population structure.

2.5 Acknowledgements

We thank Adrienne Moran-Lauter, Charlene Wight, and Andrezej Kilian for excellent technical assistance. We thank the following for contributing germplasm: Franco Asoro, Jennifer Mitchel Fetch, Weikai Yan, Stephen Harrison, Heidi Kaeppler, Ron Barnet, Mike McMullen,
Herbert Ohm, Deon Stuthman, Brian Rossnagel, Tom Fetch, Luiz Carlos Federizzi, Axel Diedericsen, Hermann Brstmayr. We thank the following for allowing us to access unpublished marker data from additional segregating oat populations: Luiz Carlos Federizzi, Itamar Nava, Steve Molnar, Kyle Gardner, Brian Rossnagel, Aaron Beattie, Eric Jackson, and Rebekah Oliver. We greatly appreciate the contributions of many additional collaborators, too numerous to list, to the germplasm collections and to the additional mapping work. Funding for this work was provided by USDA-NIFA grant number 2008-55301-18746 “Association genetics of beta-glucan metabolism to enhance oat germplasm for food and nutritional function,” Agriculture and Agri-Food Canada, and the North American Milling Association.
CHAPTER 3. MICRO-ENZYMATIC EVALUATION OF OAT (*AVENA SATIVA* L.) BETA-GLUCAN FOR HIGH-THROUGHPUT PHENOTYPING

A paper submitted to Cereal Chemistry

Mark A Newell\(^1\), Hyun Jung Kim\(^2\), Franco G Asoro\(^3\), Adrienne Moran Lauter\(^4\), Pamela J White\(^5\), M Paul Scott\(^6\), and Jean-Luc Jannink\(^7\)

Abstract

Oat (*Avena sativa* L.), a grass species grown predominantly in temperate short-season regions, has received significant attention for its positive and consistent health benefits when consumed as a whole grain food. These positive health benefits can be attributed in part to mixed-linkage (1-3, 1-4)-β-D-glucan (referred to as β-glucan), a soluble dietary fiber found in cereal endosperm cell walls. As a result of its positive impact on human health, β-glucan content has become a major focus of oat breeding operations. Unfortunately, the standard enzymatic method of measurement for oat β-glucan is costly and does not provide the high-throughput capability needed for plant breeding where thousands of samples are measured

---

\(^1\)Department of Agronomy, Iowa State University, Ames, IA 50011, USA

\(^2\)Department of Food Science and Human Nutrition, Iowa State University, Ames, IA 50011, USA

\(^3\)Department of Agronomy, Iowa State University, Ames, IA 50011, USA

\(^4\)USDA-ARS, Department of Agronomy, Iowa State University, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

\(^5\)Department of Food Science and Human Nutrition, Iowa State University, Ames, IA 50011, USA

\(^6\)Author for Correspondence, USDA-ARS, Department of Agronomy, Iowa State University, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

\(^7\)USDA-ARS, Robert W. Holley Center for Agriculture and Health, Cornell University Department of Plant Breeding and Genetics, 407 Bradfield Hall, Ithaca, NY 14853, USA
over a short period of time. The objective of this research was to test a micro-enzymatic approach for high-throughput phenotyping of oat $\beta$-glucan. Fifty North American elite lines were chosen to span the range of possible values encountered in elite oat. Evaluation for both the micro-enzymatic and standard streamlined methods for correlation, repeatability, error, and bias was conducted. Pearson and Spearman correlations ranged from 0.81 to 0.86 between the two methods. Repeatability was greater for the micro-enzymatic than for the standard streamlined approach and error was similar. Although the micro-enzymatic method did contain bias as compared to the results for the standard streamlined method, this did not substantially decrease its ability to determine $\beta$-glucan content. In addition to a substantial decrease in cost, the micro-enzymatic approach took as little as 5% of the time as compared to the streamlined method. Therefore, the micro-enzymatic method for $\beta$-glucan evaluation is an alternative method that can enhance high-throughput phenotyping in oat breeding operations.

3.1 Introduction

Mixed-linkage (1-3, 1-4)-$\beta$-D-glucan (referred to as $\beta$-glucan) is a hemicellulose that makes up about 70% of cereal endosperm cell walls (Carpita, 1996). The importance of $\beta$-glucan is most well-known in oat ($Avena sativa$ L.) and barley ($Hordeum vulgare$ L.) where breeders carry out selection in the positive and negative direction, respectively. Increased oat $\beta$-glucan has been a major target for breeding operations because of its positive and consistent health implications when oat is consumed as a whole grain. Most notably, it has been demonstrated to improve health with respect to blood pressure (Keenan et al., 2002), diabetes (Jenkins et al., 2002), cholesterol (Braaten et al., 1994), and the immune response (Estrada et al., 1997), all of which are important given the increases in human health cases related to diet over the past few decades.

Evaluation of oat $\beta$-glucan is most often done enzymatically using an approach first described by McCleary and Glennie-Holmes (1985). Since then the approach has been modified (McCleary and Codd, 1991) and sold as a streamlined mixed-linkage $\beta$-glucan assay kit (Megazyme Int., Wicklow, Ireland) that allows higher throughput capacity. Although this method works well with respect to accuracy and precision, its throughput when thousands of
samples need to be evaluated over a short time frame render it ineffective for breeding. For plant breeding, where selections are made based purely on phenotypic rank, the importance of precision outweighs that of accuracy. Thus, a method that can maintain precision and enable high-throughput phenotyping would be valuable for breeding for increased β-glucan in oat. The objectives of this research are to: 1) evaluate a micro-enzymatic method to measure oat β-glucan, 2) compare the micro-enzymatic method to the standard streamlined method, and 3) determine the repeatability and error of the two methods.

3.2 Materials and Methods

3.2.1 Micro-Enzymatic β-glucan Assay

Megazyme’s mixed-linkage β-glucan kit was used with modification to the streamlined method at one-tenth of scale. Approximately 0.5 to 3 g of groats were ground in 15 mL polycarbonate grinding vials with two stainless steel grinding balls (OPS Diagnostics, Lebanon, NJ) for three minutes at maximum speed. A Talboys High Throughput Homogenizer (Henry Troemner LLC, Thorofare, NJ) was used for grinding. Eight to 12 mg of the flour sample was weighed into 1.2 mL strip tubes arranged in a 96-well plate. Each sample was wet with 20 µL aqueous ethanol (50 % v/v) to aid dispersion followed by addition of 400 µL sodium phosphate buffer (20 mM, pH 6.5). The contents were stirred on a vortex mixer until the solute was in solution. The plate was then placed in a boiling water bath (100°C) and incubated for 15 seconds and immediately stirred on a vortex mixer until in solution. This incubation and stirring was repeated for a total of three times. Following three incubation periods of 15 seconds at 100°C, the plate was incubated at 50°C for 5 minutes. Twenty µL of lichenase (1 U) was added to each sample and stirred on a vortex mixer until in solution and incubated at 50°C for 1 hour. The plate was removed from the incubator every 10 to 15 minutes and mixed on a vortex mixer. Following the one-hour incubation, 500 µL sodium acetate buffer (200 mM, pH 4.0) was added and stirred on a vortex mixer. The plate was allowed to equilibrate at room temperature for five minutes and centrifuged at 1000 x g for 10 minutes.

Ten µL of each sample was then dispensed into two 96-well plate reader plates. To the first
plate (the measurement reaction), 10 µL β-glucosidase (0.02 U) in sodium acetate buffer (10 mM, pH 4.0) was added to each well. To the second plate, (the reaction blank), 10 µL sodium acetate buffer (50 mM, pH 4.0) was added to each sample. Both plates, the measurement reaction and reaction blank, were incubated at 50°C for 10 minutes. Three-hundred µL GOPOD reagent was added to each well and incubated at 50°C for 20 minutes. For each measurement reaction plate, one well of 10 µL D-glucose standard (10 µg) was included with 300 µL GOPOD reagent. Finally, the optical densities of the measurement reaction and reaction blank plates were determined at 510 nm using a plate reader.

The percent β-glucan on a wet weight basis was determined by the following equation:

$$\text{β-glucan} \% = \Delta A \times F \times 94 \times \frac{100}{W} \times \frac{162}{180}$$

Where: $\Delta A = \text{Absorbance of the reaction minus the absorbance of the reaction blank}$, $F = \frac{10 \mu g \text{ absorbance of } 10 \mu g \text{ of } D\text{-glucose}}{\text{ }}$, and $W = \text{Weight of the sample (ignoring moisture content)}$.

### 3.2.2 Streamlined β-glucan Assay

The streamlined β-glucan assay was implemented in accordance with Megazyme’s mixed-linkage β-glucan streamlined method (AACC Method 32-23). The percent moisture content was ignored in calculating the β-glucan content so that even comparisons could be made between the two methods. In addition to this reasoning for ignoring moisture content, the β-glucan content on a dry and wet weight basis was highly correlated (0.999) so its effect can be considered negligible. Percent moisture content for the samples analyzed covered a range of only 2 percent moisture.

### 3.2.3 Genetic Material

Four-hundred-forty-four oat lines were grown in 2009 in Ames, Iowa at the Iowa State University Agronomy Farm with two replicates. β-glucan evaluations were initially conducted as part of a large-scale genome-wide association study. Best linear unbiased predictions (BLUPs) for each line were used to choose a smaller group of 50 lines that spanned the range of β-glucan values commonly encountered in elite oat. The BLUPs (plus intercept) for the 444 lines ranged from 1.9 to 7.2% β-glucan and consisted of elite material from the United States and Canada.
The fifty lines included were chosen by sorting the lines for percent $\beta$-glucan and sampling the lines at even intervals with respect to $\beta$-glucan, excluding outliers at the tails of the distribution. Thus, the range of BLUPs for the lines resulted in a nearly uniform distribution from 2.4 to 6.3% $\beta$-glucan. The first replicate for each line was split into two samples, the first sample was evaluated using Megazyme’s streamlined method and the second was used for two replicates of the micro-enzymatic procedure.

### 3.2.4 Statistical Analysis

Each sample was evaluated once using the Megazyme streamlined and twice using the micro-enzymatic method where $\beta$-glucan was evaluated on a wet weight basis. Results for the streamlined method (ignoring dry weight) are referred to as $S_1$. An additional 20 random samples were evaluated using the streamlined method and are referred to as $S_2$. The micro-enzymatic method was applied as a completely randomized design with two replicates where plates were considered replicates. In total, testing for a significant correlation between the two methods was done at three levels. The first two correlation tests were between each micro-enzymatic replicate, or plate, and the streamlined method. These tests will be referred to as $P_1:S_1$ and $P_2:S_1$ for the first and second plate, respectively. Each plate included a D-glucose standard and for $P_1$ and $P_2$ the percent $\beta$-glucan was calculated using the D-glucose absorbance for each plate individually. The third correlation test was between the least squares means for the micro-enzymatic method across plates and the streamlined method; this is referred to as $LS:S_1$. All models were analyzed using the computer software R (R Development Core Team, 2009) using the `aov` and `lm` functions. The fixed effects model used for ANOVA was $Y = Mean + Plate + Line + error$ where $Y$ is the response, $Plate$ is the effect of plate, and $Line$ is the line effect. The mean D-glucose standard across plates was used to calculate the response values for $LS$. This was done because differences between plates should be accounted for in the plate effect. Testing for significant correlation, both Pearson and Spearman, between the micro-enzymatic method and the streamlined method was completed in R using the `cor.test` function for $P_1:S_1$, $P_2:S_1$, and $LS:S_1$.

An alternative approach that explores the relative bias when comparing measurement meth-
ods was also implemented. First explained by Altman and Bland (1983) and later reviewed by Ludbrook (2002), the method of differences evaluates two types of bias, fixed and proportional. Fixed bias is due to the change in mean value across all measurements and can be thought of as accuracy. Proportional bias is due to differences across measurements correlated to the level of analyte and is related to precision. The method plots for each paired measurement, the mean between methods versus the difference. If the mean for all points are significantly different from zero, it is an indication of fixed bias. On the other hand, if the slope of the regression is significantly different from zero, it is concluded that there is proportional bias. The method of differences was applied to three comparisons, P1:S1, P2:S1, and LS:S1.

Repeatability is relevant to comparing methods because the repeatability of two methods of measurement limit the amount of agreement that is possible (Bland and Altman, 2010). Therefore the repeatability for each method was calculated. The mixed effects model $Y = Mean + Replicate + Line + error$ where $Y$ is the response, $Replicate$ is the fixed effect of replicate, and $Line$ is the random line effect. The model was implemented twice, once for the micro-enzymatic method and once for the streamlined method. For the micro-enzymatic method, $P1$ and $P2$ were used as the first and second replicate, respectively. For the streamlined method, 20 of the lines were chosen randomly and analyzed for a second replicate, referred to as $S2$. Thus, the sample size for the two models was 50 and 20 for the micro-enzymatic and streamlined methods, respectively. The repeatability, or intra-class correlation, was calculated using the variance estimates for the random effects such that $Repeatability = \frac{Var(Line)}{Var(Line)+Var(Error)}$. The error variance is also reported. Mixed effects models were implemented in the lme4 package within the R software. Repeatability and error variance for these two comparisons are referred to as $P1:P2$ and $S1:S2$ for the micro-enzymatic and streamlined methods, respectively. In addition, the Pearson and Spearman correlations were tested for significance for $P1:P2$ and $S1:S2$.

3.3 Results

Testing for significant correlations between the two methods was done at three levels, P1:S1, P2:S1, and LS:S1. Pearson correlations were highly significant at all levels and ranged from
0.84 to 0.86 (Figure 3.1). As expected, LS:S1 showed a higher level of correlation than did P1:S1 and P2:S1, most likely because of reduced measurement error in the micro-enzymatic method due to replication. Although the high Pearson correlation between methods is important, it is the Spearman correlation that gives information about changes in selections in plant breeding that are based on phenotypic rank. Spearman correlations were highly significant and ranged from 0.81 to 0.83 for P1:S1, P2:S1, and LS:S1, respectively (Table 3.1). Since the method of correlation has been described as a poor approach for comparing methods of measurement (Ludbrook, 2002), the method of differences was also implemented. Unlike exploring correlation, the method of differences attempts to delineate fixed and proportional bias associated with the method of measurement. For all of the comparisons, the method of differences identified both fixed and proportional bias (Figure 3.2). For all three comparisons the means of the differences and the slopes of the regression lines were significantly greater than zero. A closer look at the proportional bias shows that there is a larger bias at higher levels of β-glucan content. This can be seen by the positive relationship between the mean and difference of pair-wise measurements.

Figure 3.1  β-glucan values enzymatically determined using the streamlined versus the micro-enzymatic method for P1:S1, P2:S1, and LS:S1 with 95% confidence bands. Pearson correlations for the three comparisons were significant with values of 0.85, 0.84, and 0.86, respectively. Spearman correlations for all three approaches were also significant with values of 0.81, 0.83, and 0.83, respectively.

Repeatability, calculated using the variance of the random effects for the two linear models,
Table 3.1  Summary of results for five methods of comparison including P1:S1, P2:S2, LS:S1, P1:P2, and S1:S2. P1 and P2 refer to the first and second replicates of the micro-enzymatic method, LS refers to the least squared means for P1 and P2, and S1 and S2 refer to the first and second rep of the streamlined method. Results for sample size, Pearson and Spearman correlations, repeatability, and error variance are included for each comparison.

<table>
<thead>
<tr>
<th>Method Comparison</th>
<th>P1:S1</th>
<th>P2:S1</th>
<th>LS:S1</th>
<th>P1:P2</th>
<th>S1:S2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size (n)</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>Pearson</td>
<td>0.85</td>
<td>0.84</td>
<td>0.86</td>
<td>0.94</td>
<td>0.90</td>
</tr>
<tr>
<td>Spearman</td>
<td>0.81</td>
<td>0.83</td>
<td>0.83</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Repeatability(^a)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>Error(^b)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>

\(^a\) Repeatability, or intra-class correlation, is calculated as the proportion of line variance to all sources of variation.
\(^b\) The error are the respective error variances.

was similar between the two methods. Repeatability for the streamlined (S1:S2) and micro-enzymatic (P1:P2) methods was 0.89 and 0.93, respectively. These values indicate that the repeatability is high when measured for each of the methods. More importantly, it is crucial that the methods have similar repeatability because if poor repeatability exists in one of the methods it would be expected to cause a low level of agreement between them. The error variance associated with the two methods was also similar with values of 0.06 and 0.05 for S1:S2 and P1:P2, respectively. The within method comparisons for the Pearson and Spearman correlation were similar but greater for P1:P2 than for S1:S2. Table 3.1 shows a full summary of results for all comparisons for Pearson and Spearman correlations, repeatability, and error.

### 3.4 Discussion

Before adopting a new method of measurement it is important to characterize its behavior across samples that will likely be encountered. We have presented a method for the evaluation of β-glucan for high-throughput phenotyping with high correlation with the standard method. In addition to evaluating the micro-enzymatic method based purely on the Pearson correlation, evaluations were also used to characterize change of rank and bias. Specific to plant breeding, where lines are chosen based on phenotypic rank, the Spearman correlation
Figure 3.2 Results for the method of differences to determine the types of bias in the micro-enzymatic method compared to the streamlined approach with 95% confidence bands. \( P1:S1, P2:S1, \) and \( LS:S1 \) all show fixed and proportional bias represented by the mean of values significantly greater than zero and the slope of the regression significantly greater than zero, respectively.

was implemented to address this issue. The Spearman correlation was lower than the Pearson indicating that rank changes had occurred between the two methods. The method of differences was used to characterize the bias introduced using the micro-enzymatic method compared to the standard. Results demonstrated that both fixed and proportional biases are present in the micro-enzymatic approach. Lastly, linear models were used to characterize the repeatability and error for the two methods. The relatively high and consistent repeatability across methods indicates that they perform well with respect to evaluations on the same sample. A major factor for deciding the usefulness of a method specifically for high-throughput phenotyping is the relative cost and time requirement compared to the standard approach. Based on only the cost of the kit itself, the micro-enzymatic procedure is 1/10 the cost of the streamlined method excluding the additional gain in cost for user-supplied reagents that include sodium phosphate and sodium acetate buffers. The micro-enzymatic method is most remarkable with respect to its time requirement. For grinding, including cleaning of the grinding equipment, the streamlined method takes 7 minutes per sample. The streamlined assay takes approximately 15 minutes per sample assuming 10 can be done within 2.5 hours. The micro-enzymatic method takes 45 and 15 minutes for grinding and cleaning, respectively, for 100 samples. Thus, taking
approximately 0.6 minutes per sample. For the assay itself, the micro-enzymatic method takes 3.5 hours for two plates (192 samples) or 4 hours for four plates (384 samples). Thus, the micro-enzymatic method takes in the range of 0.6 and 1.1 minutes per sample. The β-glucan assay takes approximately 0.7 and 15 minutes per sample for the micro-enzymatic and streamlined methods, respectively. Taken together, the methods take approximately 1.1 to 1.7 and 22 minutes per sample for the micro-enzymatic and streamlined methods, respectively. Thus, the micro-enzymatic procedure takes anywhere from about 5 to 8% of the time required to implement the streamlined method. Although the micro-enzymatic method does have some bias, from a practical perspective it performs quite well. The time and cost savings of the micro-enzymatic method allows for greater levels of replication than the streamlined method, potentially increasing the precision of the method. The relative cost and time requirement enable the breeder to make selections faster by reducing the cost per evaluation. Most importantly, the improvement in cost and time does not substantially decrease the methods ability to determine β-glucan content.

3.5 Acknowledgements

Funding for this work was provided by USDA-NIFA grant number 2008-55301-18746 “Association genetics of beta-glucan metabolism to enhance oat germplasm for food and nutritional function.”
CHAPTER 4. GENOME-WIDE ASSOCIATION STUDY FOR OAT (AVENA SATIVA L.) BETA-GLUCAN USING GERMLASM OF WORLDWIDE ORIGIN

A paper to be submitted to Crop Science

Mark A Newell\textsuperscript{1}, Franco G Asoro\textsuperscript{2}, M Paul Scott\textsuperscript{3}, Pamela J White\textsuperscript{4}, William D Beavis\textsuperscript{5}, and Jean-Luc Jannink\textsuperscript{6}

Abstract

Detection of quantitative trait loci (QTL) controlling complex traits followed by selection has become a common approach for selection in crop plants. The QTL are most often identified by linkage mapping using experimental F\textsubscript{2}, backcross, advanced inbred, or doubled haploid families. Although linkage mapping is a powerful approach for QTL detection, its shortcomings are numerous. An alternative approach for QTL detection is genome-wide association studies (GWAS) that use pre-existing lines such as those found in breeding programs. We explored the implementation of GWAS in oat (\textit{Avena sativa} L.) to identify QTL affecting \(\beta\)-glucan, a soluble dietary fiber with several human health benefits when consumed as a whole grain. Four-hundred-thirty-one lines of worldwide origin were tested over two years and genotyped using

\textsuperscript{1}Department of Agronomy, Iowa State University, Ames, IA 50011, USA
\textsuperscript{2}Department of Agronomy, Iowa State University, Ames, IA 50011, USA
\textsuperscript{3}USDA-ARS, Department of Agronomy, Iowa State University, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA
\textsuperscript{4}Department of Food Science and Human Nutrition, Iowa State University, Ames, IA 50011, USA
\textsuperscript{5}Department of Agronomy, Iowa State University, Ames, IA 50011, USA
\textsuperscript{6}Author for correspondence, USDA-ARS, Robert W. Holley Center for Agriculture and Health, Cornell University Department of Plant Breeding and Genetics, 407 Bradfield Hall, Ithaca, NY 14853, USA
Diversity Array Technology (DArT) markers. A mixed-model approach was used where both population structure fixed effects and pair-wise kinship random effects were included. Various mixed-models that differed with respect to population structure and kinship were tested for their ability to control for false positives. As expected, given the level of population structure previously described in oat, population structure did not play a large role in controlling for false positives. Only three markers were significantly associated with $\beta$-glucan QTL. Significant marker sequences were compared with rice and two of the three had hits localized on rice chromosome seven in a region containing the $CslF$ gene family, known to have $\beta$-glucan synthase function. Results indicate that GWAS in oat can be a successful option for QTL detection, more so with future development of higher-density markers.

4.1 Introduction

The objective of quantitative trait locus (QTL) mapping is to identify genomic regions that are associated with a specific phenotype. The identified regions, linked to a causal genetic variant, can be selected in a breeding program with the goal to improve genetic gain per unit time (Lande and Thompson, 1990). Furthermore, identification of causal variants increase our understanding of the mechanisms that affect a trait, which may in turn lead to improved selection methods. In QTL studies, experimental $F_2$, backcross, advanced inbred, or doubled haploid families are developed. Although this approach is powerful in QTL detection, the shortcomings of the approach are numerous (Jannink et al., 2001). First, the cost of generating such populations can be high, more so for advanced inbred populations and doubled haploids depending on available technologies. Second, because the populations are generated by mating two inbreds, at most two alleles at each locus are sampled. The limited diversity of such crosses limits the inference space within a breeding population. Third, the QTL mapping populations may be disconnected from the breeding process itself because inbred parents for QTL mapping are often chosen to diverge with respect to the trait of interest, rather than simply being the best individuals for breeding. Even with these shortcomings, QTL mapping does provide a powerful method of QTL detection. The high power to detect QTL linked to marker loci is due to the extent of linkage disequilibrium (LD) generated from the mating of two inbred lines and
spanning large chromosomal regions. A positive impact of such LD is the low marker density required to adequately cover the genome. Conversely, QTL positioning has low resolution such that the marker could be as much as 10-30 cM (centi-Morgans) from the causal allele (Kearsey and Farquhar, 1998).

An alternative approach to QTL mapping is genome-wide association studies (GWAS), also known as LD mapping. In contrast to QTL mapping based on bi-parental crosses, GWAS uses a sample of lines from the broader breeding population, unrelated by any specific crossing design (Zhu et al., 2008). In such studies, associations between genotype and phenotype depend on historical LD broken down by many generations of recombination. The short LD blocks that exist in such groups of lines can result in high resolution mapping of QTL. Hence, for GWAS a larger number of markers are required to assure LD between markers and causative alleles throughout the genome thus enabling fine-scale mapping. GWAS has been widely used in human genetic studies where the development of experimental populations is impossible. In contrast to the experimental populations developed for linkage mapping, a major issue facing GWAS are unknown relationships among individuals, also known as population structure, that can lead to spurious associations (Kennedy et al., 1992). To statistically control for structure, a mixed-model analysis (Yu et al., 2006) has been widely implemented.

Oat (Avena sativa L.), a grass species grown as a grain or forage crop predominantly in temperate short-season regions, poses another issue for QTL mapping. The added lack of a consensus map in oat make traditional QTL mapping far more difficult. Thus, given adequate levels of LD along with a marker system that has the ability to saturate the genome makes GWAS a suitable approach to identification of QTL in oat. Newell et al. (2010) explored genome-wide LD in oat and showed that to attain values of $r^2 = 0.2$ between markers, one marker per centi-Morgan (cM) was needed. The most comprehensive oat map available, Kanota x Ogle, is 1890 cM, thus on the order of 2000 markers would be required to reach an average LD between markers and causal alleles of 0.2. Recent advances in Diversity Array Technology (DArT) markers in oat and current single nucleotide polymorphism (SNP) development can provide such density requirements for oat.

Although oat production worldwide has been decreasing, it is still highly prized for its
positive health benefits. The health benefits associated with consuming oat as a whole grain are attributed to mixed-linkage (1-3, 1-4)-β-D-glucan (hereafter referred to as β-glucan), a hemicellulose found in cereal endosperm cell walls (Fincher, 2009). Research on the role of oat β-glucan in the human diet has been quite consistent and has been shown to improve health with respect blood pressure (Keenan et al., 2002), diabetes (Jenkins et al., 2002), cholesterol (Braaten et al., 1994), and the immune response (Estrada et al., 1997). β-glucan viscosity is a primary factor affecting the aforementioned health benefits, although the mechanisms involved are not well understood (Colleoni-Sirghie et al., 2003). Independent studies in oat and barley have demonstrated a positive relationship between viscosity and β-glucan content (Chernyshova et al., 2007; Izydorczyk et al., 2000). Thus, β-glucan content is a good target for selection in oat breeding programs.

An unintended consequence of the breeding process is the loss of genetic variants that control valuable traits (Robertson, 1960; Hill and Robertson, 1966). This is often the case for elite material where intense selection, possibly for other traits, has occurred and the useful genetic variants are lost due to fixation of the undesired allele at a locus. Thus, the identification of QTL in germplasm from worldwide origin that includes breeding lines and landraces may enable the use of genetic variants not currently found in elite varieties. The objectives of this study were to 1) conduct a GWAS to identify QTL associated with increased β-glucan content in oat germplasm of worldwide origin and, 2) determine the effects of population structure in mixed-model association analyses for oat.

### 4.2 Materials and Methods

#### 4.2.1 Genetic Material

Genetic material was requested from the National Small Grains Collection within the National Plant Germplasm System. Selection of accessions was based on two criteria, the standardized β-glucan values from the Germplasm Resources Information Network (GRIN) and the accession origin. Three data sets in the GRIN database (USDA-ARS, National Genetic Resources Program, 2010) included β-glucan information, these included oat.betaglucan.madison.07, 91,
and 95. Together these data sets included over 6000 varieties, breeding lines, and landraces of worldwide origin. Because the three data sets were measured in different years, and each set contained different lines, the values were standardized within each data set. In order to increase power for the analysis, lines were chosen that spanned the tails of the standardized \( \beta \)-glucan distribution. The second criterion for selection was based on the origin of accessions, such that lines were selected to maximize the diversity of the germplasm set. This was done to sample the array of alleles present in available oat germplasm. Approximately half of the lines selected were from the upper tail and half were from the lower tail of the distribution while taking into account the origin of the materials.

### 4.2.2 Genotypic and Phenotypic Analysis

Plants were grown under greenhouse conditions and tissue was collected from single plant accessions. Extraction of DNA was done with prescribed methods from DArT and was performed by Diversity Arrays P/L, Canberra, Australia using methods described by Tinker et al. (2009). Because genomic positions of many of the DArT markers were unknown, markers were clustered using hierarchical clustering with seven clusters and used in that order for all figures. Seven clusters were chosen based on the cluster dendrogram to adequately differentiate marker groups with similar properties. Genotyped accessions were grown as hill plots in Ames, Iowa in 2009 and 2010 in an incomplete block design. Years, replicates, and incomplete blocks were considered as fixed effects and accessions as random effects. Two replicates were grown in both 2009 and 2010 where incomplete blocks consisted of 5 x 5 hill plots. For the 2009 and 2010 season, hill plots were grown at 40 and 12 inches apart, respectively. Field checks for \( \beta \)-glucan included nine varieties and breeding lines of known \( \beta \)-glucan content. Plots were harvested, threshed, cleaned, and 0.5 to three grams of seed per hill, depending on availability, were dehulled using a compressed-air oat laboratory dehuller manufactured by Codema Inc. (Eden Prairie, MN). The field design was conserved for laboratory analysis of \( \beta \)-glucan. An enzymatic approach for evaluation of \( \beta \)-glucan content was implemented using the streamlined mixed linkage \( \beta \)-glucan kit (Megazyme Int., Wicklow, Ireland) with minor modifications. The laboratory protocol was modified to increase the throughput capability by reducing reagent
amounts by 90%, thus enabling use of a 96-well plate for analysis (Newell et al., in review). Statistical analysis for β-glucan was implemented in R (R Development Core Team, 2009) using the \texttt{lme4} package for mixed-effects models.

### 4.2.3 Data Cleaning

In order to remove possible errors and redundancies in markers and lines that may cause false associations, a data-cleaning step was implemented. This included a four-step process, all of which have been previously described as necessary steps in preparation of GWAS (Miyagawa et al., 2008). Initially, the data set consisted of 466 accessions and 1001 DArT markers. First, markers with call rates of less than 0.8 were removed; this step was implemented to remove markers that were likely to have errors. This step removed only one marker, resulting in 466 accessions and 1000 markers. Second, markers with minor allele frequency (MAF) of less than 0.01 were removed, as they do not contribute substantially to the variation in the data. This step reduced the number of markers from 1000 to 982 markers. Third, markers were merged that diverged by less than 1% across the genotyped lines, thus combining markers that were in near perfect LD. This step resulted in a matrix of 466 accessions and 796 markers. Lastly, accessions that differed by less than 1% on the markers were merged, thus removing accession redundancies. After implementation of this step, the final data set was reduced to 431 accessions and 796 markers.

### 4.2.4 Linkage Disequilibrium

One factor that can have large impacts on the power and resolution of GWAS is the extent of LD. Therefore it was explored for the data set. Of the 796 markers, 237 of the markers had known map positions in which case LD was calculated for each pair as follows: $r^2 = \left(\frac{\sigma_{XY}}{\sigma_X \sigma_Y}\right)^2$ where $\sigma_{XY}$ is the covariance between marker $X$ and marker $Y$, and $\sigma_X$ and $\sigma_Y$ are the standard deviations for marker $X$ and marker $Y$, respectively. Population structure was taken into account using singular value decomposition as previously described (Newell et al., 2010). Map distances were based on the updated Kanota x Ogle map (Tinker et al., 2009) and were used in accordance with Newell et al. (2010). The functional relationship between LD and map distance...
was determined by fitting the nonlinear model (Sved 1971) \( r^2 = \frac{1}{(1+4ad)} \), where \( d \) refers to the map distances in cM and \( a \) is an estimated regression coefficient. The parameter \( a \) can also be interpreted as the effective population size of the population to which the analysis was applied. In all, 3,968 linked marker pairs were used to fit the nonlinear model, 429 of those were with the range of zero to five cM.

### 4.2.5 Association Analysis

Association analysis to identify QTL controlling oat \( \beta \)-glucan was implemented in R using the \texttt{GWA} function with modification in the \texttt{rrBLUP} package (Endelman, 2011). The \texttt{GWA} function applies a mixed linear model that can account for both population structure and marker-based kinship, denoted \( K \), originally described by Yu et al. (2006). Marker-based kinship was calculated using the \texttt{emma.kinship} function in the \texttt{emma} package (Kang et al., 2008). Models that did not include \( K \) in the mixed-model used an identity matrix indicating no relationship between individuals.

Models accounting for differing levels of population structure fixed effects with and without \( K \) were assessed. The first model, denoted P1, included \( n_p \) number of principal components that were significantly correlated with the response variable at \( p \leq 0.01 \). Hence, \( n_p \) was chosen based purely on the number of significant axes. The second model, denoted P2, included the first \( n_p \) significant principal components, a common approach used when principal components analysis (PCA) is used to account for population structure. For both types, the number of dimensions was equal to \( n_p \), thus comparisons could be made across models. In all, six models were assessed including P1, P1K, P2, P2K, K, and a Simple model where neither P nor K was included in the model. The six models were assessed for their ability to control for type I errors by visualizing the distribution of p-values for the markers where uniformly distributed p-values indicates proper control for type I errors. The Benjamini and Hochberg (1995) false discovery rate at 0.25 was used for multiple testing of significant QTL. Two \( R^2 \) measures were used to assess the amount of variability explained for each marker. In addition to the standard measure, \( R^2 \), the likelihood-ratio based \( R^2 \), denoted \( R^2_{LR} \) was also calculated as it has been shown to be a better estimate of \( R^2 \) in GWAS (Sun et al., 2010).
4.2.6 Rice Sequence BLAST

In order to gain some information about the DArT markers that were identified as significant, the sequences were compared with rice (*Oryza sativa* L.) sequences in a three-step approach. First, a subset of candidate genes in rice were identified by searching for genes that included glucan within their functional description (Ouyang et al., 2007). Second, significant DArT sequences (Tinker et al., 2009) were compared with the entire rice genome using BLAST with an E-value threshold of $1 \times 10^{-15}$ and a hit score of greater than 500 (Ouyang et al., 2007). This level of stringency was used because it is expected there may be differences in sequence given the interspecies nature of the sequences. For markers that were previously merged during data cleaning, each marker was compared individually. Lastly, rice genomic locations for the candidate regions (step 1) and DArT sequences (step 2) were compared.

4.3 Results

4.3.1 Germplasm Selection

Selection of lines used in this study was based on two criteria, the standardized $\beta$-glucan values for three data sets from the GRIN database and the origin of lines. In total, 466 accessions were selected ranging in standardized $\beta$-glucan from 1.3 to 9.3% with a bimodal distribution. The number of lines that were classified as either high or low based on the standardized $\beta$-glucan values were 238 and 228, respectively. The lines in the distribution with lower $\beta$-glucan values ranged from 1.3 to 3.2% whereas the higher distribution ranged from 5.1 to 9.3%. Thus, selection based on this criterion was apparent in the distribution. The second criterion was selection of lines in order to increase the diversity of the germplasm set. In total, the selected accessions were from 49 countries from around the world (Figure 4.1). The majority of lines were from the United States, Turkey, Germany, and The Russian Federation with 171, 32, 28, and 27 lines, respectively. For these four countries, the lines were nearly evenly split with respect to the $\beta$-glucan classification as either high or low. In fact, for the top 14 countries that accounted for most of the lines in the set, most were evenly split between the high and low $\beta$-glucan classifications.
4.3.2 Phenotypic Analysis

Best linear unbiased predictions (BLUPs) for β-glucan ranged from -1.38 to 2.40 with an intercept for fixed effects of 4.13. Model assumptions were diagnosed by graphical representation of the residuals and the correlation between the residuals and fitted values (data not shown). As expected, due to the procedure in which the lines were selected, the distribution of β-glucan values was bimodal (Figure 4.2). Field checks ranged from -0.64 for Buff, a naked oat bred for high protein content to 2.40 for N979-5-1-22, an Iowa State University line bred for high β-glucan content. Three of the checks (HiFi, ND030288, and N979-5-1-22) had β-glucan values greater than any of the lines. The average β-glucan BLUPs for the two selection
groups according to the GRIN classifications were -0.34 and 0.34 for the low and high class, respectively. There was a highly significant correlation (0.68) between the GRIN and $\beta$-glucan BLUPs (Figure 4.3).

![Figure 4.2](image)

**Figure 4.2** Distribution of $\beta$-glucan BLUPs for the lines showing the bimodal distribution as a result of the selection process. Triangles beneath the distribution represent the $\beta$-glucan BLUPs for field checks including Buff, Excel, Winona, Cherokee, IA02130-2-2, Baker, HiFi, ND030288, and N979-5-1-22 from left to right, respectively.

### 4.3.3 Population Structure and Kinship

The level of population structure in the data set was explored to gain some insight into its possible effect on the association analysis. Principal components analysis was first implemented on the mean imputed marker data, the first three accounted for only 14.5, 6.1, and 3.7% of the total variation in the data. These low levels of variation explained, along with visualization of the principal components indicated that there is relatively weak population structure in the data as compared to other grass species such as barley (Hamblin et al., 2010). These results of
4.3.4 Linkage Disequilibrium

Previous research in oat has suggested that LD decays to an $r^2$ value of approximately 0.2 at one centi-Morgan (cM) distance (Newell et al., 2010). Of the 796 markers used in this study,
237 had known marker positions and therefore an estimate of the rate of LD decay could be determined. Based on all pair-wise estimates of LD between the 237 linked markers, the level of LD was 0.2 at 1 cM, similar to previous results (Newell et al., 2010). The estimate for $a$, the effective population size for the data set, was 94.

**4.3.5 Evaluation of P in the Mixed Model**

The effect of population structure in the mixed-model approach was tested by observation of each models ability to reduce the number of false positives. In order to assess a models ability to account for this, the distribution of observed $p$-values for the six models were plotted in the $-\log_{10}$ scale (Figure 4.4). The Null hypothesis, or expectation, follows a uniform distribution represented by a diagonal line. When marker $p$-values deviate above the Null they are significant. Thus, when there is an over-abundance of low $p$-values the distribution of $p$-values does not follow the expected uniform distribution across the majority of markers. In contrast, a model that sufficiently accounts for the number of false positives follows the expectation except for the few significant markers. Five principal components (5, 14, 25, 30, and 31) were significantly correlated to $\beta$-glucan, thus the population structure fixed effects included five dimensions. For $P1$, 5 percent of the total variation in the data was explained by the significant PCs. This is far less than for $P2$ in which the first five principal components accounted for 30 percent of the total variation. For the six models tested, $P1$, $P2$, and the Simple model did not sufficiently reduce the number of false positives. The only model that showed improvement over the Simple model in decreasing the number of false positives for these three models was $P2$, this result is most likely due to the model fixed effects accounting for a large amount of the variation in the marker data (30%). The $P1$ model did not show an improvement over the Simple model most likely for the same reason. In contrast, when $K$ was included in each of those three models, the distribution of $p$-values followed the expected uniform distribution. This indicates that the addition of $K$ in the model sufficiently accounts for relationships between individuals and effectively reduces the number of false positives. In addition, it also demonstrates the small effect of $P$ on the number of false positives when $P$ is comprised of a small and large proportion of the data variation.
Figure 4.4 Distribution of p-values for the six models included for evaluation of P in the mixed model. Axes represent the cumulative p versus the observed p in the $-\log_{10}$ scale where the dashed line represents the Null. Models that do not include K (P1, P2, and Simple) do not adequately account for false positives in contrast to P1K, P2K, and K that effectively reduce the number of false positives.

4.3.6 Association Analysis

As expected from the evaluation of P in the mixed model, models that did not include K identified a large number of significant, false positive markers. The P1, P2, and Simple models had 398, 286, and 441 significant (FDR $\leq 0.25$) markers, respectively. Given such large numbers of significant markers that are likely false positives, these models were excluded from further analyses. The numbers of significant markers were greatly reduced with the addition of K in the mixed model. Where the P1K, P2K, and K models had only three, two, and two markers significantly associated with $\beta$-glucan (Figure 4.5). Two of the significant markers were in common between the three models. These were oPt.0133 and oPt.17174/oPt.8715 where the forward slash refers to markers that were merged during data cleaning. A third marker, oPt.6825/oPt.0112, was significantly associated with $\beta$-glucan when the P1K model was im-
implemented. The three significant markers identified in the P1K model were not in LD with $r^2$ values that ranged from 0.004 to 0.031. The $R^2$ and $R^2_{LR}$ values varied around about 3% for all of the markers across models (Table 4.1).

![Manhattan plots for the eight models used for association analysis showing the scores for each marker in clustered order calculated as $-\log_{10}(p)$. Significant scores using an FDR of 0.25 are represented by bold points.](image)

**Figure 4.5**  Manhattan plots for the eight models used for association analysis showing the scores for each marker in clustered order calculated as $-\log_{10}(p)$. Significant scores using an FDR of 0.25 are represented by bold points.

### 4.3.7 Rice Sequence BLAST

As part of the three-step process for the rice sequence comparison, 130 candidate rice gene hits were identified. These covered a wide range of functions including the CslF gene family, endoglucanase, $\beta$-glucosidase, exoglucanase, and xyloglucan. The 130 hits spanned all of the rice chromosomes with most occurring on chromosomes seven, three, and six with 19, 17, and 17 hits, respectively (Figure 4.6). The fewest hits occurred on rice chromosomes 11 and 12 with two on each. In total, five oat DArT sequences were BLAST against rice because two of the significant markers had been previously merged. The DArT marker sequences included were oPt.0133, oPt.6825, oPt.0112, oPt.17174, and oPt.8715. Two of the markers, oPt.17174 and oPt.8715, resulted in no hits with the rice genome and were thus excluded from further evaluation. The remaining three markers, oPt.0112, oPt.6825, and oPt.0133, had 1, 3, and
Table 4.1 Results for the significant markers for the P1K, P2K and K models including the score, $R^2$, $R^2_{LR}$ and the FDR $q$-value obtained using the Benjamini and Hochberg method for multiple testing. The FDR cutoff for significance was 0.25. Marker names separated by a forward slash represent merged markers as a result of the data cleaning steps.

<table>
<thead>
<tr>
<th>Marker Name</th>
<th>Model</th>
<th>oPt.0133</th>
<th>oPt.6825/oPt.0112</th>
<th>oPt.17174/oPt.8715</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P1K, P2K, K</td>
<td>3.33, 3.69, 3.48</td>
<td>3.15</td>
<td>3.09, 3.29, 3.41</td>
</tr>
<tr>
<td>Score (-log$_{10}(p)$)</td>
<td>P1K</td>
<td>2.7, 3.2, 3.0</td>
<td>2.6</td>
<td>2.5, 2.8, 2.9</td>
</tr>
<tr>
<td>$R^2$ (%)</td>
<td>P1K, P2K, K</td>
<td>2.8, 3.1, 2.9</td>
<td>2.6</td>
<td>2.5, 2.7, 2.8</td>
</tr>
<tr>
<td>$R^2_{LR}$ (%)</td>
<td></td>
<td>0.22, 0.16, 0.16</td>
<td>0.22</td>
<td>0.22, 0.20, 0.16</td>
</tr>
<tr>
<td>$q$-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

30 hits respectively (Figure 4.6). The hit for oPt.0112 was located on rice chromosome seven in position 34110 and encoded an integral membrane protein $DUF6$ containing protein. The three hits for oPt.6825 included the same hit for oPt.0112 in addition to two more hits on chromosome seven in positions 34320 and 34340 encoding nodulin and a conserved hypothetical protein, respectively. Hits for oPt.0133 occurred on eight of the 12 rice chromosomes with most occurring on chromosomes four, five, eight, and seven. Although there were many gene hits in rice for oPt.0133, all of the hits were protein kinases, protein kinase receptors, protein receptor-like kinases, $SHR5$ receptor-like kinases, or $DUF26$ kinases. Interestingly, oPt.0133 occurred in clusters of three or more for 24 out of the 30 hits. One of those clusters, on rice chromosome seven, occurred in tandem with the four hits of oPt.0112 and oPt.6825. Results for the aligned BLAST DArT sequences with the rice candidates are shown in Figure 4.6. One specific region contained hits for all three DArT sequences, rice chromosome seven from positions 34000 to 37000. The rice candidates surrounding position 35500 include the glucan endo-1,3-beta-glucosidase precursor and those surrounding position 36700 encode the $CslF$ genes, a gene family with $\beta$-glucan synthase function (Fincher, 2009).

4.4 Discussion

Numerous research studies have been implemented for GWAS using the mixed-model approach that accounts for population structure and kinship initially described by Yu et al. (2006).
Figure 4.6 Arrangement of β-glucan rice candidates and oat DArT sequences on the rice genome. Rows represent the rice chromosome number and a plus symbol represents precise positions of the sequences. The entire length of each chromosome is not shown; high and low regions with no DArT sequence hits are not included.

For this study, we evaluated the inclusion of population structure fixed effects in the model. We found that including $P$, as principal components, in the model did not substantially decrease the number of false positives. Also, the number of false positives decreased as the amount of variation that the principal components accounted for increased. Similar results were found for simulated data where the $K$ model performed as well or better than models including population structure (Bradbury et al., 2011). Their study found this result to be consistent across varying numbers of QTL and heritability estimates. Similarly, the initial publication by Yu et al. (2006) found that the including population structure showed an improvement over not including it only for traits highly correlated to population structure. This could partially explain the results in our study where the small influence of $P$ in the mixed-model are reminiscent of the low levels of population structure that exist in oat. In contrast to these results, Stich et al. (2008) found that including $P$ in the mixed-model showed greater improvement in controlling for false positives than just including $K$. Although, as pointed out in the manuscript, this
could be a likely cause of the high levels of population structure in the data set.

In addition to the evaluation of $P$ in the mixed-model, three markers were identified to be associated with increased $\beta$-glucan content. Two of which were in common for the three models that included $K$ and one that was only significant in the $P1K$ model. Interestingly, when $P1K$ was implemented as opposed to $P2K$ there was another significant marker identified. Two previous studies have identified QTL associated with increased $\beta$-glucan content in oat. One of these studies (Kianian et al., 2000) developed two recombinant inbred line populations and identified two regions on linkage groups 11 and 14 that were consistent across populations and all environments. Two of the significant markers identified in this study had known map positions (Tinker et al., 2009) and are located on linkage group eight. An unfortunate result of this study is that only three QTL were identified as being significantly associated with increased $\beta$-glucan content. A possible reason for this result is that the decay of LD (0.2 and 1 cM) is too rapid for the marker density to capture much of the QTL variation that exists. Hence, higher marker densities would likely be a solution to this issue and would enable higher QTL resolution.

The three identified QTL were also compared with rice sequences in a three-step process that should enable the presumed QTL to be matched by location to potential rice candidates. The three-step approach was implemented because it is likely that the DArT sequences, having not been identified for function previously, may not have perfect alignment with a specific rice gene but may fall in a region with likely candidates. DArT sequences that fall within regions of likely candidates may support the notion that the DArT marker is truly associated with $\beta$-glucan content. Three of the markers had positive hits in the rice genome, two of which were all located on rice chromosome seven. The third marker also had hits on rice chromosome seven in addition to hits distributed across the rice genome with sequence similarity to protein kinases. An interesting result of this comparative approach was that all three of the markers clustered in a short segment on chromosome seven. Rice candidates in this region included the $CslF$ gene family, previously identified in rice (Burton et al., 2006) and barley (Burton et al., 2008) to have $\beta$-glucan synthase function. Additionally, because one of these sequences was only significant using $P1K$, and still localized to this common region in rice, it indicates that
using significantly correlated PCs to the response is a good method to account for population structure. Although the QTL sequences do not fall within the CslF gene sequences, this does suggest that the identified markers are true $\beta$-glucan QTL. To date, $\beta$-glucan QTL have not been identified in oat within proximity to the CslF gene family in other grass species. The results presented indicate that the lines in the NPGS could possess valuable alleles for $\beta$-glucan content not found in elite oat.

### 4.5 Acknowledgements

Funding for this work was provided by USDA-NIFA grant number 2008-55301-18746 “Association genetics of beta-glucan metabolism to enhance oat germplasm for food and nutritional function.”
CHAPTER 5. DETERMINING THE NUMBER OF CLUSTERS FOR HIGH-DIMENSIONAL DATA WITH APPLICATION TO EXPLORING CROP POPULATION STRUCTURE

A paper submitted to The Annals of Applied Statistics

Mark A Newell\(^1\), Dianne Cook\(^2\), Heike Hofmann\(^3\), and Jean-Luc Jannink\(^4\)

Abstract

A first step in exploring population structure in crop plants and other organisms is to define the number of clusters or 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.

\(^1\)Author for Correspondence, Department of Agronomy, Iowa State University, Ames, IA 50011, USA
\(^2\)Department of Statistics, Iowa State University, Ames, IA 50011, USA
\(^3\)Department of Statistics, Iowa State University, Ames, IA 50011, USA
\(^4\)USDA-ARS, Robert W. Holley Center for Agriculture and Health, Cornell University Department of Plant Breeding and Genetics, 407 Bradfield Hall, Ithaca, NY 14853, USA
5.1 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. 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. 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 subpopulations. In fact, although current clustering methods, such as $k$-means and hierarchical are quite useful, they do very little to address the practical question of how many clusters exist (Fraley and Raftery, 2003). 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 et al., 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 addi-
tion 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 means 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 the exponential growth of hypervolume as a function of dimension, thus making high-dimensional data sets problematic. In contrast, Murtagh (2009) determined that in very high-dimensional space there is a simplification of structure, demonstrating that the distances within and between clusters become tighter 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). We note that, similar to other data types, genetic data is usually asymptotic where the high dimension, low sample size (HDLSS) situation occurs. Hall et al. (2005) found that low-dimensional projections of such data sets where the number of dimensions $d \rightarrow \infty$ while the number of observations $n$ is fixed tend to lie at vertices of a regular simplex, in agreement with Murtagh (2009). (Note that the supplementary material for this paper contains video of higher dimensional views of the plant breeding data that also support this simplified structure). HDLSS data can pose a challenge when implementing principal component analysis
(PCA) because the covariance matrix is not of full rank. This leads to a strong inconsistency in the eigenvectors in which case the added variation obscures the underlying structure of the covariance matrix (Jung and Marron, 2009). We note the importance to keep in mind the ensuing issues when implementing PCA for HDLSS data sets. In particular, clustering methods that do not work well for high-dimensional data such as model-based approaches 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 the empirical data sets. The computer software GENOME (Liang et al., 2007), a coalescent-based whole genome simulator, offers just this by simulating sequences backwards 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.

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 just 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 5.2 describes the algorithm for choosing clusters. Section 5.3 describes the simulated and empirical data sets used to validate the algorithm. Section 5.4 describes the results. Supplementary material 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.

5.2 Algorithm for Choosing the Number of Clusters

The algorithm to determine the number of clusters has four steps; bootstrap sampling, clustering, calculation of cluster validity statistic, and the computation of a permutation test for significance. Hubert’s gamma statistic (Halkidi et al., 2001), available in the R package fpc (Hennig, 2011), is the cluster validity statistic of choice, chosen heuristically from many criterion because it simply seemed to work better than others, and calculate easily across clustering methods. It is also on a standard scale which makes comparison between methods simpler. For consistency, matrices are denoted in bold typeface with the subscript representing the number of rows and columns, respectively. Let $X_{n \times p}$ ($n$ 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 GENOME 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 5.1 (left). The user sets a maximal number of clusters, $k_{\text{max}}$, based on prior knowledge of a maximum. The steps are then as follows:

1. **Bootstrapping**: A number of bootstrap samples, $b$, is drawn at random from the rows of $X$ with replacement. The resulting matrix is denoted as $X_{n \times p}^* i$ for $i = 1, \ldots, b$.

2. **Cluster analysis**: Three methods of cluster analysis are implemented for $1, 2, \ldots, k_{\text{max}}$
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 $Y_{n \times 1}, Y_{n \times 2}, Y_{n \times 3}, \ldots, Y_{n \times k_{\text{max}}}$ where the number of clusters is set to $k$. Thus the number of principal components is equal to the number of clusters.

(b) Hierarchical (hclust) clustering is applied to the Manhattan distance matrix $D_{n \times n}$ 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 $X_{n \times p}^*$ with the number of clusters set to $k$.

3. Cluster validity: For each 1, 2, ..., $k_{\text{max}}$ 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, ..., $k_{\text{max}}$ 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 $k-1:k_{\text{max}}$. 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 5.1 (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.
Figure 5.1 (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.

5.3 Data

5.3.1 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 5.2 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.

![Figure 5.2](image)

Figure 5.2  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.

### 5.3.2 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., 2010). 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, 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., 2011). 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 5.3). 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 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 5.2 and 5.3 demonstrate how the low-dimensional projections from PCA are quite similar between the simulated and empirical sets, 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 5.1.

5.4 Results

5.4.1 Simulated Data

Results for 100 simulated data sets at each cluster - migration rate combination are summarized in Table 5.2. 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 simulated. 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.
Figure 5.3 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 5.1 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.

<table>
<thead>
<tr>
<th>Source (Name)</th>
<th>Crop</th>
<th>Origins</th>
<th>Line Types</th>
<th>Dimensions</th>
<th>Marker Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newell et al. (2010)</td>
<td>oat</td>
<td>World</td>
<td>Varieties, breeding lines, landraces</td>
<td>1205 x 403</td>
<td>DArT</td>
</tr>
<tr>
<td>(newell2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asoro et al. (2011)</td>
<td>oat</td>
<td>North American</td>
<td>Elite cultivars</td>
<td>446 x 1005</td>
<td>DArT</td>
</tr>
<tr>
<td>(asoro2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tinker et al. (2009)</td>
<td>oat</td>
<td>World</td>
<td>Varieties, breeding lines, landraces</td>
<td>198 x 1958</td>
<td>DArT</td>
</tr>
<tr>
<td>(tinker2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamblin et al. (2010)</td>
<td>barley</td>
<td>US</td>
<td>Elite cultivars</td>
<td>1816 x 1416</td>
<td>SNP</td>
</tr>
<tr>
<td>(hamblin2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SNP</td>
</tr>
<tr>
<td>Zhang et al. (2009)</td>
<td>barley</td>
<td>Canada</td>
<td>Cultivars, breeding lines</td>
<td>169 x 971</td>
<td>DArT</td>
</tr>
<tr>
<td>(zhang2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chao et al. (2010)</td>
<td>wheat</td>
<td>US, Mexico</td>
<td>Spring/winter wheat cultivars</td>
<td>478 x 219</td>
<td>SNP</td>
</tr>
<tr>
<td>(chao2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 5.4 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 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 5.4). 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 5.5 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 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
Table 5.2  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.

<table>
<thead>
<tr>
<th>True No</th>
<th>Migration Rate</th>
<th>Mean Est No</th>
<th>True Hubert’s Gamma</th>
<th>Hubert’s Gamma</th>
<th>Proportion Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00005</td>
<td>6.09</td>
<td>1.00</td>
<td>0.48</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>5.37</td>
<td>1.00</td>
<td>0.47</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>5.79</td>
<td>1.00</td>
<td>0.48</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0.00005</td>
<td>2.09</td>
<td>0.69</td>
<td>0.73</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>2.47</td>
<td>0.49</td>
<td>0.55</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>3.46</td>
<td>0.36</td>
<td>0.48</td>
<td>0.38</td>
</tr>
<tr>
<td>3</td>
<td>0.00005</td>
<td>3.00</td>
<td>0.77</td>
<td>0.79</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>3.17</td>
<td>0.58</td>
<td>0.61</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>3.50</td>
<td>0.46</td>
<td>0.52</td>
<td>0.56</td>
</tr>
<tr>
<td>4</td>
<td>0.00005</td>
<td>4.00</td>
<td>0.78</td>
<td>0.79</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>4.12</td>
<td>0.59</td>
<td>0.62</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>4.49</td>
<td>0.46</td>
<td>0.51</td>
<td>0.48</td>
</tr>
<tr>
<td>5</td>
<td>0.00005</td>
<td>4.99</td>
<td>0.76</td>
<td>0.78</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>5.09</td>
<td>0.59</td>
<td>0.63</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>5.40</td>
<td>0.47</td>
<td>0.51</td>
<td>0.41</td>
</tr>
<tr>
<td>6</td>
<td>0.00005</td>
<td>6.02</td>
<td>0.75</td>
<td>0.77</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>6.04</td>
<td>0.58</td>
<td>0.62</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>6.34</td>
<td>0.46</td>
<td>0.51</td>
<td>0.48</td>
</tr>
<tr>
<td>9</td>
<td>0.00005</td>
<td>8.97</td>
<td>0.72</td>
<td>0.75</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>9.06</td>
<td>0.55</td>
<td>0.59</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>8.53</td>
<td>0.43</td>
<td>0.46</td>
<td>0.51</td>
</tr>
<tr>
<td>12</td>
<td>0.00005</td>
<td>11.97</td>
<td>0.68</td>
<td>0.71</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>0.00010</td>
<td>11.36</td>
<td>0.51</td>
<td>0.55</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>0.00015</td>
<td>9.20</td>
<td>0.41</td>
<td>0.43</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Figure 5.4  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.
Figure 5.5  Diagnostic plot of Hubert’s gamma for varying number of clusters solutions 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, \textit{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.
gamma is positively correlated with the proportion of times the correct number of clusters was chosen with a value of 0.84 (Table 5.2). 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.

5.4.2 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 5.3 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 are also represented as the number of clusters versus the Hubert’s gamma statistic in Figure 5.6. 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. Zhang2010 has a unique 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 5.3). 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 bootstrap samples, in which case this would be sufficient for this data set.
Table 5.3  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.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>No of Clusters, Method</th>
<th>Previous Results</th>
<th>Previous Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>newell2010</td>
<td>4. kmeans (0.481)</td>
<td>4. kmeans (0.481)</td>
<td>5, kmeans (0.526)</td>
</tr>
<tr>
<td>asoro2011</td>
<td>3, kmeans (0.434)</td>
<td>4, kmeans (0.424)</td>
<td>5, kmeans (0.431)</td>
</tr>
<tr>
<td>tinker2009</td>
<td>1, - (1)</td>
<td>1, - (1)</td>
<td>1, - (1)</td>
</tr>
<tr>
<td>hamblin2010</td>
<td>6, hclust (0.816)</td>
<td>6, hclust (0.816)</td>
<td>6, hclust (0.816)</td>
</tr>
<tr>
<td>zhang2009</td>
<td>2. kmeans (0.786)</td>
<td>2, kmeans (0.786)</td>
<td>2, kmeans (0.786)</td>
</tr>
<tr>
<td>chao2010</td>
<td>3, kmeans (0.581)</td>
<td>4, hclust (0.579)</td>
<td>4, hclust (0.579)</td>
</tr>
</tbody>
</table>

Figure 5.6  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 $p < 0.01$. 
Newell2010 required 200 bootstrap samples to reach equilibrium with respect to the numbers 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 5.3 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. (2011) 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 5.3). 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. Lastly, 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 lines and the 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, the Hubert’s gamma statistics for each are shown simultaneously in Figure 5.7. The variation in the true Hubert’s gamma for the simulated data sets at each cluster - migration rate combination covers a range of about 0.1, 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. Lastly, 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.
Figure 5.7 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.

5.5 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 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.

In agreement with two previous studies (Hall et al., 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 set 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.

5.6 Acknowledgements

We would like to thank Eduard Akhunov, Franco Asoro, Francois Belzile, and Nick Tinker for providing empirical data sets for testing. Funding for this work was provided by USDA-NIFA grant number 2008-55301-18746 “Association genetics of beta-glucan metabolism to enhance oat germplasm for food and nutritional function.” Authors Cook and Hofmann’s work on this paper has been partly supported by the National Science Foundation grant DMS0706949.
CHAPTER 6. OUTREACH EDUCATION - TECHNIQUES TO ENGAGE MIDDLE SCHOOL STUDENTS

A paper submitted to The Journal of Curriculum and Pedagogy

Mark A Newell

Abstract

Scientists present research on a regular basis to a wide variety of audiences, but rarely have the opportunity to effectively present to a younger audience. Engaging the middle school student can be a challenging task that may require techniques not usually appropriate to scientific audiences. More recently, scientists are participating in outreach opportunities involving middle school students given its high priority in receiving external funding. Science Days, an event organized for middle school students, exposes youth to a variety of topics and allows them to have direct contact with scientists. The objective of the three-day event was to present students my research on plant breeding, specifically related to oat quality improvement. This manuscript discusses the techniques used during Science Days with a major focus on techniques that were most appropriate to increase student engagement. The expected outcome of this manuscript is to provide basic techniques for other scientists, including graduate students, to engage middle school students in future outreach opportunities.

1Author for Correspondence, Department of Agronomy, Iowa State University, Ames, IA 50011, USA
6.1 Introduction

Scientists routinely give presentations and seminars during outreach activities, but generally do not have the opportunity to present their research in a meaningful way to younger audiences. Within the National Science Foundation Graduate K-12 (NSF-GK-12) fellowship, PhD candidates are given the unique opportunity to develop the skills necessary to communicate with young students both in and out of the classroom. The recent trend for increasing outreach programs such as GK-12 has stemmed from the drive to strengthen science education in the United States on an international level. The GK-12 program is funded by the National Science Foundation as part of a nationwide initiative to place scientists into classrooms. This initiative has been shown to be a useful tool to increase the quality of science education (Baumgartner et al., 2006) in the United States. The timing of such programs is also critical. It has been found that student interest and success in science starts to decline in middle school and continues thereafter (Catsambis, 1995). The Iowa State University GK-12 project, Symbi (Symbi, 2011), targets middle school students in typical science classrooms over the course of one school year. Each PhD candidate is paired with a teacher and is immersed in a classroom to leverage the fellows research experiences in order to develop innovative science activities to engage the young students.

The benefits of a program such as GK-12 are three-fold: 1) research has suggested that interactions with a scientist on a regular basis, engaging the students with authentic, hands-on activities, generate new views of science and scientists, 2) teachers benefit by learning new science content and new ways to teach the material, and enjoy the collegial support, and 3) the fellows gain teaching skills, acquire a greater understanding of education, and improve their communication skills with a general audience (Laursen et al., 2007). Communication from the scientist to the non-scientist, including the general public, can be a novel and challenging task for scientists (Leshner, 2007). Mathews et al. (2005) reported in their qualitative study with geneticists that many scientists felt they did not have the necessary experience or skills to effectively communicate with a non-science audience. A simple question, such as What do you do in your research? can be a difficult question to answer when communicating with audiences
outside of the scientists research community. However, this is fundamentally an important question that any scientist should be able to answer with an optimum final outcome - a good explanation of the research and why it is relevant to the person asking the question.

At the end of the school year, Symbi has an outreach event termed Science Days. The one-day event took place at three different schools over the course of three weeks where scientists and fellows were encouraged to discuss their own research with the students. Over the course of the day, groups of six to twelve 6th, 7th, and 8th grade students visited each scientist’s table for ten minutes. The desired outcomes of this activity were to: 1) expose the students to a variety of science fields, and 2) increase student interests in science. As a PhD candidate in plant breeding, my objective for Science Days was to present my research in oat (Avena sativa L.) quality improvement, specifically for human consumption. The audience consisted of middle school students at three schools from the Des Moines Public School District. My involvement in the Symbi program for the 11 months prior to Science Days provided me with the background experience needed to appropriately gauge the audience and apply the strategies learned in the classroom to this unique outreach opportunity. This manuscript discusses the successes of various techniques that I used at the Science Days event with major focus on student engagement.

6.2 Materials and Methods

The major objective of Science Days was to present my research on oat quality improvement for human consumption, specifically for increasing dietary fiber (beta-glucan). It has been well established that oat dietary fiber has many positive health benefits that include reducing the risk of heart disease, helping control diabetes, and lowering cholesterol. Because of the health benefits associated with oat consumption, the title of the middle-school presentation was Super Food for Super Human Strength. The materials used in my presentation included an immature (green) oat plant, a mature oat plant, hulled oats, de-hulled oats (groats) and hulls, oatmeal, a Cheerios® box, and a computer. The general flow of the presentation included four major sections as follows:
1. *Discovering the oat plant:* Based on my previous 11-months in the classroom, it was safe to assume that the students would not know anything about an oat plant so the presentation started with, Does anyone know what kind of plant this is? holding an oat plant in my hand. As I presumed, most students did not know what it was, in fact only two out of the 302 students who visited me over the three-week period were correct. I then proceeded to discuss the green oat plant in that photosynthesis was still occurring and how it was still producing energy from the sun. From there, I turned the students attention to the mature oat plant and allowed them to touch the two plants at different maturities and make their own comparisons. I then asked them What part of the plant do we eat? and showed them the hulled oats and explained the purpose of the hull, to protect the seed. The process of de-hulling was then explained showing both the de-hulled oat groats (the portion that is eaten) along with hulls. Lastly, I asked them What kinds of food do you eat that contain oats? showing them Cheerios® and oatmeal, specifically pointing out the first ingredient in Cheerios® – whole grain oats. At this point I made an effort to explain to the students what the term whole grain oats meant by showing them the de-hulled oats. I also showed them the similarities between the de-hulled oats and oatmeal, which was easily noticeable when they were side-by-side.

2. *Establishing relevance to the audience:* To assist the students in understanding the relevancy of the research, I asked them What is the leading cause of death in the United States? In some cases, the students did answer correctly with heart disease. The students or myself also mentioned some of the other major causes of death that were diet related (e.g. some types of cancer, stroke, and diabetes). In most cases the students personally knew someone affected by one of these diseases. In order to make the connection between oats and positive health benefits, I then read straight from the Cheerios® box that consuming oat dietary fiber may reduce the risk of heart disease. This was probably the most important section for the students. If this connection between oats and health had not been made explicit, the students would have been at risk of leaving the presentation without the personal relevance necessary to form a long-term understanding.
3. Explanation of the research: After establishing relevance, I referred back to the initial oat plant and explained my research. Specifically talking about using a collection of oat plants from around the world, finding which ones have the most dietary fiber in the grain and therefore healthier, and then mating them to make oat plants with improved dietary fiber. At this point, I gave a simple definition of plant breeding, improving a plant for a specific purpose. I used other examples to demonstrate the process of breeding as well, including improvement of cotton for clothing, improvement of chicken for human consumption, and tree improvement to enhance the quality of wood used to make the very basketball court we were standing on. The students were quite amazed by all the examples of what breeders did and the range of species that were bred for specific purposes.

4. Applying technology to answer questions: The last section focused on a major part of my research, statistics. I first posed the question to the students, What is statistics? Most of the students were not really sure. In order to relate the unfamiliar concept to something the students already knew, I asked them if they watched sports on television and if they had ever heard of the term player stats. Since most of the students knew what this was, I defined statistics as simply the collection and interpretation of data. I also used a simple example of collecting the height of all the students in the gymnasium and comparing the height of males and females. I then showed them a scatterplot of about 1200 data points representing different oat plants in my research and asked them if they had ever made a plot with more than 20 points, in most cases the answer was no. The idea was to show the students a multi-dimensional tour of my data set on the computer using GGobi (Swayne et al., 2003). I first related it to a movie seen in 2-dimensional space compared to the detail seen in a 3-dimensional movie. The students were amazed with the tour (Asimov, 1985; Buja et al., 2005; Cook et al., 2006). The main reason for the data visualization tour was to demonstrate how I used computers in my research. This element of the presentation was definitely attractive for some of the students who were more interested in electronics.
The four sections outlined above are a basic description of the presentation; in some cases all of the sections were not completed. Due to the fact that I am not a teacher trying to get through content, I could remain flexible and discuss with the students the specific topics related to my research that were more interesting to them at the time. If I finished the presentation before the 10 minutes were over, I asked the students about their interests in science.

### 6.3 Techniques to Engage Middle School Students

Throughout the three-day Science Days event, various techniques enabled me to engage the students in my research.

1. *Be personable:* Middle school students, just like adults, prefer to interact with someone whom is pleasant to be around. During my interactions with students, simply smiling, making jokes, laughing, and being enthusiastic showed the students that scientists are people also someone they could and would want to be like in their futures. A common and required theme of outreach programs in engineering to engage students cites the importance of engaged role models that are able to excite and educate K-12 students (Jeffers et al., 2004). In agreement, being personable and maintaining a good attitude was one of the most valuable techniques used to engage the students. Although this seems trivial, it can be a difficult task to maintain for two reasons: 1) the length of time interacting with the students, and 2) the repetitive nature of the day. During the Science Days events, the same presentation was given approximately 20 times throughout the day. I quickly discovered that if my enthusiasm dropped during the afternoon presentations, students were less engaged in the material and more quiet in the discussion.

2. *Breakdown barriers:* Two types of barriers seemed to lessen student engagement physical and social. The major physical barrier in my situation was the table provided for presentation materials such as my computer and oat specimens. By standing out in front of the table I was able to create more of a discussion situation with the students; I was talking with them rather than lecturing at them. Additionally, this technique lent itself to facilitating discussion by allowing the students to stand around me in a circle rather
than in a line similar to the classroom. This developed a situation of cooperative learning within the students and between the students and myself. Another major physical barrier between the students and myself was that fact that I was taller than most of the students. Since I was out in front of the table, I was able to lower myself and in some cases, where only few students were at my table, kneel down to speak with the students more easily. This technique also helped me to re-focus the students on the discussion if neighboring scientist-presenters were creating a distraction.

As much as scientists would rather not believe, social barriers do exist between the scientist and students. The best technique I used to overcome this barrier was to keep it as informal as possible introducing myself by first name and, when possible, using their first names. Most importantly, this included not using conventional salutations such as Dr, Mrs, Ms, and Mr similar to what a teacher or professor uses in the classroom. By using these salutations it placed an authoritative barrier between the student and scientist. For this situation, my role was not a teacher, but was a scientist whose main task was to expose the students to my research. The students were much more comfortable to speak. They seemed to view me as more an equal and less an authority figure; therefore, leading to a good discussion. This approach has also been used by popular scientist and comedian Bill Nye who goes by Bill Nye the Science Guy, who has had unprecedented success in student outreach programs.

3. **Make the relevance of your research clear**: One of the most important techniques to engage the students was to design the presentation in such a way as to clearly demonstrate the relevance of the research. Research is always relevant to the scientist conducting the research, but extra thought may by required to define why the research is relevant to a middle school student. In my case, the leading causes of death in the United States were put on a personal level with the students by making them think about people they knew who were affected. This was followed by a clear objective of the research relating it to the leading causes of death. If the student does not leave with a clear understanding of how the research personally affects them, then the student will soon forget the research
all together. This can be a difficult technique to implement along with the language used for explanation, but a considerable amount of time should be appropriated as such. In line with this technique, previous outreach partnerships have suggested that reaching the personal interests of students is a major factor in increasing student engagement (Scott et al., 2011).

4. *Stay on the science:* Because the major objective of Science Days was to engage the students in my research, it was very important not to lose the science by oversimplifying the presentation so that the research was lost or forgotten. It is tempting to design an easy and fun hands-on activity for the students, but it is essential to make sure that the activity is closely tied to what the scientist wants them to take home. If the students leave the demonstration table unsure of why they were doing an activity and what they were supposed to have learned, then clearly the activity was overly simplistic and merely entertainment. To design a presentation that is engaging as well as educational can require more thought and effort on the part of the scientist. However, it is critical to engage the student in a meaningful way. Obviously, the minor details of the research a scientist conducts at the doctoral-level is much more advanced than the majority of middle school students can understand; therefore, finding the big picture content that pushes them to think without oversimplification is a challenge. For me, finding this optimum came with past experience with middle school students and was improved throughout the Science Days event.

5. *Make them think:* In the ideal situation, the students leave the presentation thinking about what was discussed. The best way to do this was to open a dialog by asking them questions, and thus guide students to discover the correct direction for themselves. This type of teaching style is helpful for the students because it encourages them to generate explanation on their own, a hallmark of guided inquiry learning (To-im and Ruenwangsa, 2009). Making them think also kept them alert. During my experience, I asked a lot of different questions including: 1) What is a super food? 2) What kind of plant is this? 3) What is the leading cause of death in the United States? 4) Do you know anyone
who has been affected by any of these causes? 5) What part of the plant do humans consume? and 6) What do you eat that contains oats? All of these questions led the students to specific topics that I, the scientist, would like to discuss. I discovered that the range of questions changed over the three-week period as I became more experienced with presenting the material. The types of questions asked were much in-line with the previous technique, stay on the science, where finding the correct questions for the specific audience was crucial.

6. **Show full processes:** Before the experience, I believed that it was vital for the students to see the oat plant from beginning to its end product. I was confident that this would further solidify the relevance of my research to the students by showing them parts of it that were recognizable to them, for example, Cheerios®. The entire process for my research included: 1) an immature oat plant, similar to a field plant at mid-season, 2) a mature plant typical of a plant at the end of the season ready for harvest, 3) hulled oats, 4) de-hulled oats plus hulls, and 5) Cheerios® and oatmeal. By capturing the entire process from the raw material to what was more recognizable to the students, they could relate to the research as a whole. This can be a challenge with some research, but even pictures or descriptions can assist others to visualize a process.

7. **Always hold props:** The props in my case (e.g. whole plants, seeds, oatmeal, etc.) helped to capture attention by serving as something else the students could look at and establish some level of curiosity. I found that when discussing the research (section three above) having the plant in hand and using it as a visual engaged the students better than not holding anything. After discussing the prop, it could then be handed off to the students to hold, giving them another avenue to connect to the research. This may have been because having something interesting keeps the audiences eyes straight ahead with little chance for distraction. In any case, this was a large part in keeping the students engaged throughout the presentation always having something to show them.
6.4 Closing Remarks

Engaging middle school students on a research topic such as the improvement of oat quality with respect to human consumption can be a daunting task. However, by implementing the above techniques, I could engage the students who most likely had no interest or awareness in the topic beforehand. Similar to most fields of science, recently plant breeders have emphasized the importance of outreach events (Stuber and Hancock, 2008). In order to draw positive outcomes from outreach opportunities such as exposing students to the field of plant breeding and demonstrating its importance to the public as a whole, the plant breeder must have some ability to engage the audience (Hancock and Stuber, 2008). Likewise, the National Air and Space Museum (NASA, 1996) has placed a high value on its outreach program for both public awareness and K-12 outreach. As the importance of outreach continues to grow, it is critical to provide scientists with techniques that will result in positive outcomes from such events.

6.5 Acknowledgements

I would like to thank Lindsey Long for her unprecedented hard work and dedication towards the Symbi project that made it a success, my advisors William Beavis and Paul Scott for their continued support, and all others involved with Symbi. Funding for Symbi, the Iowa State GK-12, was provided by the National Science Foundation grant DGE-0947929. The research was also partially funded by USDA-NIFA grant number 2008-55301-18746 “Association genetics of beta-glucan metabolism to enhance oat germplasm for food and nutritional function.”
CHAPTER 7. GENERAL CONCLUSIONS

7.1 Conclusion

This dissertation provides initial steps towards selection for genotype in oat to improve its nutritional value. In any research that attempts to identify QTL controlling complex traits, the extent of LD affects the resolution of genomic regions that are identified. Faster and slower decay of LD result in relatively higher and lower resolution, respectively, given appropriate marker density. For GWAS, where the lines are not developed experimentally, population structure can also have a large impact on the results, contributing to an excess of markers falsely associated with the phenotype. Chapter 2 addresses these issues in a world collection of oat germplasm specifically discussing its implications for GWAS. The results from Chapter 2 in conjunction with a high-throughput approach for β-glucan evaluation (Chapter 3) enable the implementation of GWAS to identify QTL controlling β-glucan content (Chapter 4). As described, Chapter 2 explored the level of population structure in oat. This was done by assigning the lines to groups using cluster analysis and exploring different aspects of the identified clusters. Although this approach was sufficient, a method that could identify the number of clusters for such data sets using various clustering methods and choosing the most appropriate on some criterion would be optimum. Chapter 5 proposes an algorithm for identifying the number of clusters in multivariate data with application to genotypic data in crop plants. The proposed algorithm uses bootstrapping, three methods of clustering, and a common criterion for selection of the optimum number of clusters and method of clustering. A final and important step of
research that is often overlooked is establishing its relevance to the public. Two common approaches to accomplish this are through extension and outreach activities. Chapter 6 presents a case-study for an outreach activity specifically focused on techniques to engage middle school students. The benefits such activities are numerous for students and researchers. For students it has the potential to increase their interests in plant breeding and more broadly, the STEM (Science, Technology, Engineering, and Math) fields. For researchers, it increases their ability to communicate to non-scientific audiences. Lastly, it generates new views of science and scientists within the general public. As a whole, this dissertation completes a circle of research that states the objectives, designs and implements research to address the stated objectives, and finally disseminates the results through outreach education. The implications of such research benefit the oat community by establishing methods with potential to increase the response to selection, the public by direct nutritional improvement of a food source, and young students by generating their interest and awareness in the field of plant breeding.


