Predicting crown rot resistance in wheat using a complex backcross population

L. B. Zweep (s1267752)

First supervisor: Dr. M. Malosetti
Second supervisor: Dr. F. A. van Eeuwijk

MASTER’S THESIS
Defended on November 1, 2017

Specialization: Statistical Science

STATISTICAL SCIENCE FOR THE LIFE AND BEHAVIOURAL SCIENCES
# Contents

1 Abstract 3  
2 Introduction 4  
3 Methods 6  
   3.1 Data 6  
      3.1.1 Pedigree structure 6  
      3.1.2 Phenotypic data 7  
      3.1.3 Genomic data 10  
   3.2 Statistical analysis 10  
      3.2.1 GWAS SNP selection 10  
      3.2.2 Genomic prediction models 11  
      3.2.3 Allele frequencies 12  
   3.3 Software 13  
4 Results 14  
   4.1 Phenotypic data 14  
   4.2 Exploratory GWAS analysis and comparison of kinship matrices 14  
   4.3 Genomic prediction 16  
   4.4 Allele frequencies 18  
5 Discussion 20
List of Figures

1  Pedigree structure of CRI2-0428 ...................................................... 8
2  Manhattan plot with selected markers ........................................... 15
3  Comparison of GWAS outcomes with different kinships ............... 15
4  Scatterplots of predicted CRI2 lines against the measured values of the CRI2 lines, based on the different models and the different training sets .......................... 17

List of Tables

1  Browning Assessment Scale ........................................................... 7
2  Genetic variance, residual variance, heritability and model of the experimental set-up for the different glasshouse trials ............................................................. 9
3  Correlations between the glasshouse trials with the CRI0 lines ......... 14
4  Correlations between the glasshouse trials with the CRI2 lines ........ 14
5  Correlations of predicted values for different models with measured values of the CRI2 lines .......................... 16
6  The linear model with selected markers and their changes in allele frequency (AF) .................. 18
1 Abstract

Fusarium crown rot is a damaging disease frequently found in wheat, which is mostly caused by *Fusarium pseudograminearum*. Crown rot resistance is quantitatively inherited complex trait caused by many small effects loci. Lines with crown rot resistance have been identified that are interesting potential sources of favorable resistance alleles. In order to make use of this resistance in agriculture, lines need to be developed which contain a combination of high crown rot resistance and other agro-economically important traits, like high yield. In this study, these lines have been bred by crossing multiple crown rot resistant donor lines with elite lines. The lines are mostly selected by phenotype, but scoring of crown rot severeness is difficult. In this study genomic prediction of crown rot resistance was explored to partially or totally replace phenotyping in the selection process. Two different generations in a complex wheat population (an early generation CRI0 lines and a later generation CRI2 lines) were used to make genomic prediction models and to validate these models. A genomic best linear unbiased prediction (G-BLUP) model, a linear model based on a set of selected markers and a Gaussian kernel model were trained on the early generation (CRI0 lines) and validated on the later generation (CRI2 lines). Prediction based on early generation information was disappointingly low, so suggesting that phenotyping cannot be fully avoided at the later generation. The alternative of partial phenotyping at the late generation yielded more encouraging results. The Gaussian kernel model and the G-BLUP model yielded a predicted ability of about 0.41. While still further researcher is needed, the results so far imply genomic prediction within a population could be useful to select highly crown rot resistant lines.
2 Introduction

Fusarium crown rot (CR) is a widespread and damaging disease occurring in multiple agriculturally important species, like wheat and barley. The species *Fusarium pseudograminearum* is most commonly found to cause CR (Akinsanmi et al., 2004). CR occurs in all regions of Australia, but has the largest effect in the Northern region, where it can potentially cause 22% yield loss per year (Murray and Brennan, 2009). In Australia the costs of this pathogen are estimated to be around 79 million Australian dollars per year. There are managing practises to reduce CR infection, like burning stubble or rotating crops, but these techniques are not ideal, as burning stubble decreases soil moisture and rotating crops does not guarantee the *F. pseudograminearum* is gone before the next cycle starts (Liu and Ogbonnaya, 2015).

New techniques have been sought to decrease the prevalence of CR in agriculture. The breeding of crown rot resistant lines has become one of the main objectives in this search (Wildermuth et al., 2001). Several accessions of wheat have been found to have some degree of CR resistance. The aim is to breed lines with a combination of high agro-economical value and CR resistance. Elite lines can be crossed with crown rot resistant lines in order to obtain such lines. Lines can be selected on phenotype, but it would be more cost efficient to use the genotypic information in selection and could even increase higher selection precision. The lines do not all have to be phenotyped in big replicated trials, which is often more expensive than genotyping plants. The precision could be enhanced, because phenotyping with high precision is difficult. Phenotyping would still be needed, but the scale of testing could be reduced.

Both genotypic and phenotypic selection yield some difficulties. The first difficulty is the phenotyping of CR infection. Research on how to phenotype CR accurately is still being conducted (Wallwork et al., 2004; Li et al., 2008). At this moment, the stem browning (SB) score is often used as measure of crown rot severeness. The second difficulty is finding quantitative trait loci (QTL) on the genome. The detection of QTL is an important part in developing crown rot resistant lines. If the CR resistance of lines can be predicted by measuring the genotype and selecting on QTLs, the lines do not all need to be grown and phenotyped.

Crown rot resistance is a quantitative trait which is regulated by multiple loci on the genome (Bovill et al., 2006). QTLs are already identified on several places on the wheat genome. The chromosomes containing QTLs for CR resistance are found to be 1A, 1B, 1D, 2A, 2B, 2D, 3B, 4B, 4D, 5D, 6D and 7A (Martin et al., 2015; Collard et al., 2005; Bovill et al., 2010; Collard et al., 2006; Poole et al., 2012; Zheng et al., 2014; Ma et al., 2012; Li et al., 2010; Bovill et al., 2006). The QTL on 3B has the largest effect and was detected in multiple studies.

QTL detection and genomic prediction are often done on standard populations, like two-parent recombinant inbred lines (RIL) (Poole et al., 2012), double haploid lines (DH) (Zheng et al., 2014) and
backcrossed lines (BC) (Martin et al., 2013). These populations are limited by the crossing of just two parents. With two parents, the gene pool for finding QTLs is limited to segregating to two alleles per locus. In many recent studies, multiple parental lines have been crossed to create multiparental populations (Milner et al., 2016; Mackay et al., 2014; Verbyla et al., 2014). Within multiparental populations, the phenotypic and genotypic variation are bigger, which increases the genomic diversity, which increases the chance of detecting QTLs. The effects of the alleles can also be estimated in a more diverse background (Solberg Woods, 2014). In this study, we used data from a multiparental breeding population developed by crossing different crown rot resistant lines with a number of elite lines. This population was used to explore how well CR resistance can be predicted by marker information using genomic prediction approaches.

Genomic predictions can help in breeding and selecting CR resistant lines. Instead of growing and phenotyping all plants in a population, the population can be genotyped and the lines with highest predicted CR resistance can be selected for further breeding. Genomic prediction can be based on certain QTLs or on the whole genome. When building a prediction model within a population, the relation between the lines should be modeled, because the lines are related by pedigree. This relation can be either based on pedigree information or on the genetic information. Two techniques were explored in this study: QTL-based prediction based on the results of a genome wide association study (GWAS) and genetic best linear unbiased prediction (G-BLUP). In these techniques, the kinship is estimated from the genetic information. The GWAS is used to identify regions on the genome which are associated with the CR resistance to be used as predictors. The G-BLUP model and Gaussian kernel are used for genomic prediction based on the full genome without a pre-selection of predictors.

The aim of this study is to explore how well genetic prediction works to predict CR resistance between two generations in a complex multiparental cross. Can genomic prediction be used as a tool in breeding of crown rot resistant lines with high argo-economical value?
3 Methods

3.1 Data

The data used for the analysis has been obtained as part of a crown rot project at the University of Queensland (UQ) led by Dr Mark Dieters. The project started in 2010. A set of 14 founder varieties were selected, of which eight were crown rot resistant donor lines and six were elite lines with high agro-economical value. A complex cross was made and two separate generations within the cross were genotyped and phenotyped, these were called the CRI0 lines and the CRI2 lines.

3.1.1 Pedigree structure

The elite lines include 11 varieties which were or still are used in commercial breeding. The six varieties that were selected at the start of the project are Sunguard, Ventura (UQ Agenta Pure), Crusader (Longreach Plant Breeding), Scout (Longreach Plant Breeding), Gregory (UQ Agenta Pure), Gladius (UQ RSP Pure) and Magenta (UQ RSP Pure). Additional varieties were later introduced for the development of the CRI2 lines: Mace (UQ RSP Pure), Wallup, Sunguard, Spitfire and Emu Rock. The donor lines contain Kukri/Janz#37 (K/J#41, SARDI), Frame/Sentinel#41 (F/S#41, SARDI), E17 (CSIRO), L2-94 (CSIRO), LRC2009-119 (DEEDI), LRC2009-012 (DEEDI), AUS33735 (AWCC) and AUS33802 (AWCC).

The eight donor lines were crossed with the six elite lines, resulting in 48 crosses (F₁). The F₁ were backcrossed with their respective elite parent, resulting in 48 F₂BC₁. The F₂BC₁ were crossed with F₂BC₁ within their elite parental background, but combining different donor parental backgrounds. This F₃ generation was selected on crown rot resistance and selfed for two generation (F₃:4). Finally, seed was increased for the selected lines resulting in the CRI0-xxxx lines (F₅), where xxxx contains the identity of the line.

The CRI2 lines were developed using the CRI0 lines. The CRI0 lines were backcrossed either with their elite parent or a matched elite variety. CRI0 lines with Gregory and Scout background were backcrossed. CRI0 lines with Ventura, Crusader and Magenta background were crossed with Sunguard, Spitfire and Mace respectively. CRI0 lines with Gladius background were crossed with both Wallup and Emu Rock. The F₁ were crossed with F₁ within the same elite background and F₂ were crossed with F₂, also within the same elite background. The resulting F₃ descended from one or two matched elite parents and four to eight donor lines. This is comparable with an eight-way cross or multi-parent advanced generation.

---

1 South Australian Research and Development Institute
2 Commonwealth Scientific and Industrial Research Organisation
3 Department of Employment, Economic Development and Innovation
4 Australian Winter Cereals Collection
inter-cross (MAGIC) population. The lines were selected on crown rot resistance and selfed for two
generations ($f_{4.5}$), resulting in the CRI2-xxxx-1 lines. These CRI2 lines were grown in a glasshouse and
genotyped. As example, the pedigree of one of the lines (CRI2-0428) is shown using Helium (Shaw et al.,
2014) in Figure 1.

3.1.2 Phenotypic data

From the CRI0 377 lines were selected and phenotyped for crown rot, yield and rust in glasshouse trials
between 2013 and 2015. From the CRI2 575 lines were selected and measurements were done on the same
traits in glasshouse trials between 2015 and 2016. The seeds were sown in pots and after germination,
the seeds were inoculated with $F. pseudograminearum$. In total, seven different trials were conducted:
four trials with the CRI0 and parental lines and three trials with the CRI2 lines and ten parental lines
as checks. The trials each had a different experimental lay-out.

Stem browning scores (SB) were noted for each plant as score for the effect of crown rot infection
on a plant. Originally a 0-4 point scale was used to assess the crown rot severeness, but in the March
2016 and June 2016 CRI2 trials, 5 was added to the scale to indicate death due to crown rot infection
(Table 1). This new scale point was treated as score of 4 in the prediction, because it is the most severe
score on the data set that was used as predictor.

In April 2013 the first glasshouse trial was conducted on a set of 375 CRI0 lines and nine parental
lines at the UQ (GH8.1). One line was planted in two different pots in a complete randomized design with
about fifteen seeds per pot. The second trial started in 2013 (GH2) on 376 CRI0 lines and 18 parental
lines. Inoculated plants were grown on 18 benches. The lines were randomly allocated within pots and
the pots were arranged in benches of 72 pots. There were four replicates of the CRI0 lines. During the
trial diseases and pests were observed in parts of the glasshouse.

In June 2014 a glasshouse trial with a subset of 62 of the CRI0 lines and 20 parental lines was conducted
(GH8.2). Trays with ten by eight cells were used instead of pots. This experiment was replicated four

<table>
<thead>
<tr>
<th>Score</th>
<th>Appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>no crown rot symptoms</td>
</tr>
<tr>
<td>0.5</td>
<td>slight browning</td>
</tr>
<tr>
<td>1</td>
<td>browning to first node</td>
</tr>
<tr>
<td>1.5</td>
<td>if first node is blackened</td>
</tr>
<tr>
<td>2</td>
<td>browning to second node</td>
</tr>
<tr>
<td>2.5</td>
<td>if second node is blackened</td>
</tr>
<tr>
<td>3</td>
<td>more than first 2 nodes</td>
</tr>
<tr>
<td>3.5</td>
<td>extensive browning, but seed in head</td>
</tr>
<tr>
<td>4</td>
<td>white head</td>
</tr>
<tr>
<td>5</td>
<td>dead from crown rot (only in trials mar16 and jun16)</td>
</tr>
</tbody>
</table>
Figure 1: Visualization of the pedigree structure for the CRI2 line CRI2-0428. The resulting line contains the genes of five different donors and the two matched elites Ventura and Sunguard.
times with nine trays. August 2014 the last CRI0 crown rot trial started (GH8.3), using pots instead of trays. Within every pot, ten seeds were sown from one line and every line was sown in two different pots. The pots were spread out over four different benches. The lines were divided into three different batches.

For the CRI2 lines, three trials were conducted in December 2015 (dec15) with 555 CRI2 lines, March 2016 (mar16) with 561 CRI2 lines and June 2016 (jun16) with 69 CRI2 lines. These trials had similar experimental set-ups. Over multiple rooms, benches with 96 pots per bench were used. Every pot contained three different lines. All experiments contained four replicates per CRI2 line and six to eight replicates of the ten parental lines. In the jun16 trial only one room was used.

For all traits the best linear unbiased estimators (BLUEs) were computed based on the experimental set-up. The correlations between the different trials were calculated to get insight in how well the phenotypes of the lines corresponded with each other over the different trials.

In order to make predictions, the phenotype trials were combined by calculating the weighted mean over the BLUEs. The BLUEs were obtained for every trial using the R-package lsmeans (Lenth, 2016), by taking the marginal means over the fixed line effects in a linear model. The same models, but with lines as random effects, were used to estimate the genetic variance and heritability. The models with lines as random effects and their variances are shown in Table 2.

The models contained an effect for the lines, $g_l$, and effects for experimental set-up. In the models $r_i$ is the replicate, $c_j$ is the room, $b_k$ is the bench, $p_m$ is the pot, $t_n$ is the tray and $a_o$ is the batch of the $h^{th}$ individual. Underlined letters indicate random effects.

The BLUEs were weighted by the inverse of the BLUE error variance and averaged per line. This resulted in a weighted mean phenotype for the CRI0 lines and a weighted mean phenotype for the CRI2 lines. Because the BLUEs were calculated, the spatial information was already taken into account in these SB scores. These phenotypes were used in the prediction and marker selection.

<table>
<thead>
<tr>
<th>Trial</th>
<th>$V_G$</th>
<th>$V_E$</th>
<th>$H^2$</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH8.1</td>
<td>0.405</td>
<td>0.497</td>
<td>0.62</td>
<td>$y_{lh} = g_j + e_{lh}$</td>
</tr>
<tr>
<td>GH2</td>
<td>0.010</td>
<td>0.186</td>
<td>0.19</td>
<td>$y_{ikmlh} = r_i + r_{bh} + r_{bp} + g_l + e_{ikmlh}$</td>
</tr>
<tr>
<td>GH8.2</td>
<td>0.091</td>
<td>0.639</td>
<td>0.37</td>
<td>$y_{nlh} = r_i + r_{tn} + g_l + e_{nlh}$</td>
</tr>
<tr>
<td>GH8.3</td>
<td>0.085</td>
<td>0.437</td>
<td>0.33</td>
<td>$y_{ikol} = r_i + r_{rb} + r_{rbp} + g_l + e_{ikol}$</td>
</tr>
<tr>
<td>dec15</td>
<td>0.060</td>
<td>0.810</td>
<td>0.23</td>
<td>$y_{jkmlh} = c_j + c_{bjk} + c_{bp} + g_l + e_{jkmlh}$</td>
</tr>
<tr>
<td>mar16</td>
<td>0.091</td>
<td>0.425</td>
<td>0.46</td>
<td>$y_{jknlh} = c_j + c_{bjk} + c_{bp} + g_l + e_{jknlh}$</td>
</tr>
<tr>
<td>jun16</td>
<td>0.064</td>
<td>0.237</td>
<td>0.52</td>
<td>$y_{kmnlh} = b_k + p_{km} + g_l + e_{kmnlh}$</td>
</tr>
</tbody>
</table>
3.1.3 Genomic data

The CRI0 lines, a selection of CRI2 lines and the parental lines were genotyped using a 90 000 single nucleotide polymorphism (SNP) array. The array was analysed and the SNP loci with a call rate over 40% and a minor allele frequency (MAF) of at least 5% were selected. Heterozygotes were assigned as uncalled (NC) genotype. The raw genotype data contained allelic states AA and BB, NC and missing values (NA). NC values were treated as NA values.

The total missingness in the raw genotype data was 11.5%. Some lines and markers contained over 80% missingness. To increase power, markers and lines with a missingness over 30% were filtered out of the data. The cleaned genotype data had 7.4% missingness. The genotype data contained 933 lines of which 341 CRI0 lines, 567 CRI2 lines and 25 other lines including the parental lines. In total 25 092 SNPs were used, of which 19 337 SNPs were mapped on the wheat genome by Wang et al. (2014).

3.2 Statistical analysis

Three approaches were used to investigate how well CRI0 lines (n = 353) could be used to predict CRI2 lines (n = 564), to find out whether the older generation could be used in prediction of the newer generation. The first approach was based on GWAS to identify SNPs with high predictive power. The SNPs significantly associated with the phenotype were used as predictor variables in a linear model. The second prediction method was G-BLUP (Piepho et al., 2008; Piepho, 2009; Meuwissen et al., 2001), which is based on all SNPs. The last approach was a Gaussian kernel model. A cross-validation within the CRI0 lines was conducted to explore and compare the predictive power of the three models. Finally, the prediction models were build on the different training sets to predict the CRI2 lines.

3.2.1 GWAS SNP selection

With genome-wide association study (GWAS), SNPs with high association with crown rot resistance were selected to be part of a linear prediction model. For GWAS, a linear mixed model was fitted with one marker and tested. The p-values were transformed by taking the negative base ten logarithm (−log(p)). Higher −log(p) values were more correlated with the outcome and were selected for the linear model.

The linear model used in this approach is

\[ y = X\beta + Zg + S\tau + \epsilon \]  

(1)

where \( X \) is the design matrix for the fixed effects, other than the SNP effect, \( \beta \) contains the fixed effects parameters other than the marker effect, including an intercept, \( \tau \) is the marker effect parameter for the \( i \)th SNP, \( S \) is the SNP matrix with the SNP genotype, \( Z \) is the design matrix for the genetic
effects, \( g \) are the random genetics effects and \( \varepsilon \) are the random residuals with \( \text{Var}(\varepsilon) = \sigma^2_\varepsilon \). The genetic effects are based on the kinship matrix by \( \text{Var}(g) = K\sigma^2 \). The kinship matrix was built by using the R-package synbreed, which computes the kinship as realized relatedness between individuals (VanRaden, 2008). For the calculation of the kinship, missing marker information was imputed at random, because the calculations require no missing values.

To estimate the kinship matrix, either markers used in the GWAS (mapped markers) or markers not used in GWAS (unmapped markers) were used. The set of markers was thinned, removing highly correlated markers. The objective of this thinning was to avoid putting more weight on the positions with a high density of SNPs. Markers with a correlation of 0.5 or higher were discarded in a stepwise procedure. Lastly, the chromosome specific kinship matrices were estimated based on all markers except those on that specific chromosome. This was done to prevent the tested SNP effect from being included in the kinship matrix. Three different kinships were eventually built and compared: a kinship with all SNPs, a chromosome specific kinship with only mapped markers and a chromosome specific kinship with only unmapped markers.

Selecting the SNPs for the kinship was done by using a stepwise algorithm, in which the correlation between the SNPs was calculated per chromosome. One of the two SNPs with the highest correlation was removed from the SNP data and the correlation was calculated again. This procedure was repeated until there were no SNPs with a correlation higher than 0.5 left within a chromosome. To increase computation speed, large SNP sets were evaluated step by step, by taking the first 50 SNPs, removing the highly correlating SNPs, adding a new set of 50 SNPs and repeating this.

The GWAS was performed with the mapped marker chromosome specific kinship and the unmapped marker chromosome specific kinship. In each step the SNP with the highest \(-\log(p)\) value was selected and added to the model as a cofactor (\(X\)). Selected SNPs were added in the new GWAS model until no SNPs with a \(-\log(p)\) higher than 2.5 were found. The threshold of 2.5 was chosen to make a rough division between SNPs with a big effect and SNPs with a small effect. These selected SNPs were combined in a linear model for prediction.

### 3.2.2 Genomic prediction models

The genomic prediction models were based on the classical G-BLUP model, which is similar to a GWAS model, but without a term for a fixed SNP effect. The kinship was calculated by including all SNPs. The classical G-BLUP model only uses additive effects. The model is as follows

\[
y = \mu + Zg + \varepsilon
\]  
(2)
where \( \mu \) is the overall mean, \( Z \) is the design matrix for the random effects, \( g \) is the BLUP solution for the genetic values, with the kinship matrix \( \text{Var}(g) = K\sigma_g \), \( \varepsilon \) are the residuals with \( \text{Var}(\varepsilon) = \sigma^2 \).

As an extension of the G-BLUP model, a genomic prediction model with Gaussian kernel was used to predict the CRI2 lines from the CRI0 lines. The Gaussian kernel model uses a Euclidean distance matrix.

\[
G_{ij} = e^{(K_{ij}/\theta)^2} V_g
\]

in which \( K \) is the Euclidean distance matrix, \( \theta \) is a scale parameter, chosen by calculating the log-likelihood over different \( \theta \) and \( V_g \) is the genetic variance as calculated by restricted maximum likelihood (REML). The Euclidean distance matrix \( (K) \) was calculated based on the full marker data set with the distances between all lines. The prediction model with Gaussian kernel allows for more complex relations between the marker effect than an additive model (Gianola and van Kaam, 2008).

To estimate the genomic prediction within the CRI0 lines, a five-fold cross-validation was conducted. The G-BLUP model, the Gaussian kernel model and the linear model with SNPs selected from GWAS were used to calculate the predicted values and to correlate these with the measured values. In cross-validation the linear model with selected SNPs was tested for a model with threshold for the \(-\log(p)\) value of 2.5.

Genomic prediction with the three models was done using three different training sets: the CRI0 lines, a subset of the CRI2 lines (\( n = 200 \)) and a combination of CRI0 lines and the subset of CRI2 lines (\( n = 553 \)). Prediction with the CRI0 lines was explored to find out whether the phenotype could be predicted from past data. The prediction on the CRI2 lines was explored to find out whether the phenotype could be predicted based on a part of the CRI2 lines, to decrease the number of lines needed to be phenotyped. The combination of both sets was used to find out whether genomic prediction would perform better when information from past data was included. The correlation between the predicted values and the measured values was calculated.

### 3.2.3 Allele frequencies

The allele frequencies for the SNPs in the CRI0 lines and the CRI2 lines were calculated to keep track of changes due to the backcross and selection cycles. The SNP file was split into the part with only the CRI0 lines and only the CRI2 lines to calculate the allele frequencies. The difference between the allele frequencies in CRI0 lines and CRI2 lines was calculated and compared by a permutation test. In this test two populations with the same sample sizes as the CRI0 and CRI2 lines were randomly sampled thousand times and the median change in allele frequency was calculated. An empirical 95% confidence interval was computed this way. The median absolute difference between the CRI2 lines and the CRI0 lines was
calculated and compared with the 95% confidence interval. The allele frequencies of the SNPs selected in the GWAS were investigated to find out whether the change in allele frequency could be associated with the phenotype.

3.3 Software

The data preparation and analysis were mainly done in the open source software environment R, using the integrated development environment RStudio. Different software packages were used. For coding the genotype file and calculating the kinship matrices the R-package synbreed was used (Wimmer et al., 2012). The GWAS and G-BLUP analyses were done with the R-package rrBLUP (Endelman, 2011).
4 Results

4.1 Phenotypic data

Table 2 (section 3.1.2) contains the variances and heritabilities in the different glasshouse trials. The genetic variance was comparatively low as compared to the residual variance in most trials. GH8.1 had the highest genetic variance in comparison with the residual variance, resulting in a high heritability (0.62). The mar16 trial had intermediate heritability (0.46). The other trials had a lower heritability. There was quite some difference between the trials in terms of the heritability estimates. The correlations between the phenotypes in the different trials were very low and even negative between some trials (Table 3 and Table 4). The CRI0 lines gave very different outcomes over the different trials. GH8.1 and GH8.2 correlated the most (0.213). GH8.2 correlated negatively with GH8.3. Overall, the correlations were low. Between the trials with the CRI2 lines the correlation between mar16 and jun16 was higher (0.441). The dec15 trial did not correlate positively with the other two trials.

Table 3: Correlations between the glasshouse trials with the CRI0 lines

<table>
<thead>
<tr>
<th></th>
<th>GH2</th>
<th>GH8.1</th>
<th>GH8.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH8.1</td>
<td>-0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GH8.2</td>
<td>0.023</td>
<td>0.213</td>
<td></td>
</tr>
<tr>
<td>GH8.3</td>
<td>0.099</td>
<td>0.090</td>
<td>-0.189</td>
</tr>
</tbody>
</table>

Table 4: Correlations between the glasshouse trials with the CRI2 lines

<table>
<thead>
<tr>
<th></th>
<th>mar16</th>
<th>dec15</th>
</tr>
</thead>
<tbody>
<tr>
<td>dec15</td>
<td>0.092</td>
<td></td>
</tr>
<tr>
<td>jun16</td>
<td>0.441</td>
<td>-0.202</td>
</tr>
</tbody>
</table>

4.2 Exploratory GWAS analysis and comparison of kinship matrices

Figure 2 shows the Manhattan plot for the GWAS with the $-\log(p)$ scores against the position on the chromosome. This GWAS was performed using the unmapped specific kinship. The Manhattan plot shows peaks above $-\log(p)$ of three at chromosomes 1A, 2A, 2B, 3A, 3B, 4A, 5A, 5B, 6A, 7A and 7B. These peaks are places on the genome that were significantly associated with the SB scores and indicate possible QTL positions. The red points are the SNPs that were chosen for the linear prediction model. Fourteen SNPs were selected based on the $-\log(p)$ threshold of 2.5. The selected SNPs were all significantly associated with the SB scores. In the GWAS with the mapped chromosome specific kinship the same markers were selected as with the unmapped kinship.

The GWAS was done two more times with the mapped chromosome specific kinship and the standard kinship. Figure 3 shows the scatter plots of the obtained $-\log(p)$ values for the different kinship building
techniques. The red line indicates a one-to-one relation, which was expected if kinship matrices did not influence the detection power of correlating SNPs in a GWAS. The $-\log(p)$ values were higher for the unmapped specific kinship and the mapped specific kinship as compared to the standard kinship. Figure 3C shows the unmapped specific kinship yielded slightly higher $-\log(p)$ values than the mapped specific kinship, but this difference was smaller than the difference with the standard kinship.

Figure 2: Manhattan plot of the GWAS results with unmapped chromosome specific markers kinship. The red dots are the selected markers.

Figure 3: Comparison of GWAS outcomes for the standard kinship, unmapped specific marker kinship and mapped specific marker kinship. The red line is expected if there is no difference between the GWAS with different kinships.
4.3 Genomic prediction

In the cross-validation the three models were explored on their genomic prediction within the CRI0 lines. The SNP model with boundary value of 2.5 in GWAS yielded a correlation of 0.22 between the predicted and measured values. The G-BLUP model yielded a correlation of 0.29 and the Gaussian prediction model yielded a correlation of 0.28.

The three different predictions models were compared on the three different training sets. The linear model with the SNPs from GWAS, the Gaussian kernel model and the G-BLUP model were based on the CRI0 lines, a subset of the CRI2 lines and a combination of the CRI0 and CRI2 lines subset. The correlations of the predicted values and the measured values of the CRI2 lines are summarized in Table 5 and the scatterplots in Figure 4.

The linear model, based on the CRI0 lines, was build with the selected SNPs from the GWAS as fixed effects (Table 6), with 14 SNPs which had a $-\log(p)$ above 2.5. The predictions from the linear model on the CRI2 lines had a correlation of 0.174 with the measured values. The proportion of variance explained was 0.03. The model based on the CRI2 lines subset had a correlation of 0.271 and included ten SNPs. The model based on the combination of CRI2 lines had a correlation of 0.315 and included 17 SNPs. The scatterplots of the linear models are shown in Figure 4A, B and C.

The correlation between the predictions from the G-BLUP model, based on the CRI0 lines and the measured values was 0.047 (Figure 4D). The G-BLUP model based on a training set from the CRI2 lines performed better (Figure 4E). A correlation of 0.403 was found between the measured and predicted values of the CRI2 test set ($r = 0.403$, $p < 0.001$). The combination of the CRI0 lines and the CRI2 training set (Figure 4F) yielded a similar result ($r = 0.395$, $p < 0.001$), with both a proportion of variance explained of 0.17.

The Gaussian kernel model performed slightly better, but not much different from the G-BLUP model. The correlation with predictions from the CRI0 lines was very low (0.050, Figure 4G), while the models trained on the CRI2 subset and the combination of CRI0 and a subset of CRI2 lines yielded a higher prediction: 0.408 and 0.409 respectively (Figure 4H and I).

Table 5: Correlations of predicted values for different models with measured values of the CRI2 lines.

<table>
<thead>
<tr>
<th></th>
<th>CRI0</th>
<th>CRI2 subset</th>
<th>CRI0 and CRI2 subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear model</td>
<td>0.1741</td>
<td>0.2717</td>
<td>0.3153</td>
</tr>
<tr>
<td>G-BLUP</td>
<td>0.0473</td>
<td>0.4027</td>
<td>0.3945</td>
</tr>
<tr>
<td>Gaussian kernel</td>
<td>0.0497</td>
<td>0.4083</td>
<td>0.4088</td>
</tr>
</tbody>
</table>
Figure 4: Scatterplots of predicted CRI2 lines against the measured values of the CRI2 lines, based on the different models and the different training sets. A, B and C are the genomic predictions from the ten SNPs selected in GWAS. D, E and F are the genomic predictions from the G-BLUP model. G, H and I are the genomic predictions from the Gaussian kernel model.
4.4 Allele frequencies

The allele frequencies between random populations and the CRI0 and CRI2 lines were compared. The upper limit of the empirical 95% confidence interval from the permutation was 0.025 and the median difference in allele frequency for the CRI0 and CRI2 lines was 0.081, which is outside of the empirical confidence interval. The CRI0 lines and CRI2 lines differ in allele frequencies. Figure 5 shows the absolute change in allele frequency is plotted against the position on the chromosome. The allele frequency differences were present across the genome. A change in the overall allele frequency was expected in a backcross population under selection as this one.

The allele frequencies and changes in allele frequencies between CRI0 and CRI2 for the selected SNPs are shown in Table 6. A positive change means the allele frequency has increased from the CRI0 lines to the CRI2 lines. The sign of these changes in allele frequencies can be related to the effects from the linear model in the prediction. The marker Kukri_c22349_754 had a negative effect estimate, indicating a positive effect on crown rot resistance. The change in allele frequency for this model was positive, indicating an increase of the favorable allele. The marker Tdurum_contig12711_252, on the other hand, had a positive effect, indicating a negative effect on crown rot resistance. The allele frequency is also positive, showing an increase of this unfavorable allele.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Marker name</th>
<th>Effect estimate</th>
<th>CRI0</th>
<th>CRI2</th>
<th>Change in AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>Kukri_c22349_754</td>
<td>-0.060</td>
<td>0.051</td>
<td>0.116</td>
<td>0.065</td>
</tr>
<tr>
<td>1D</td>
<td>Kukri_c28164_739</td>
<td>-0.048</td>
<td>0.278</td>
<td>0.511</td>
<td>0.233</td>
</tr>
<tr>
<td>2A</td>
<td>Tdurum_contig12711_252</td>
<td>0.049</td>
<td>0.292</td>
<td>0.302</td>
<td>0.011</td>
</tr>
<tr>
<td>3B</td>
<td>wsnp_CAP8_c1419_836650</td>
<td>0.071</td>
<td>0.026</td>
<td>0.104</td>
<td>0.078</td>
</tr>
<tr>
<td>4A</td>
<td>BobWhite_c32313_688</td>
<td>-0.065</td>
<td>0.258</td>
<td>0.083</td>
<td>-0.175</td>
</tr>
<tr>
<td>4A</td>
<td>GENE-0689_658</td>
<td>-0.070</td>
<td>0.298</td>
<td>0.110</td>
<td>-0.188</td>
</tr>
<tr>
<td>4A</td>
<td>IAAV2216</td>
<td>0.035</td>
<td>0.508</td>
<td>0.455</td>
<td>-0.053</td>
</tr>
<tr>
<td>5A</td>
<td>Ex_c21738_1159</td>
<td>0.057</td>
<td>0.535</td>
<td>0.383</td>
<td>-0.152</td>
</tr>
<tr>
<td>5A</td>
<td>Excalibur_c22411_801</td>
<td>0.045</td>
<td>0.275</td>
<td>0.492</td>
<td>0.217</td>
</tr>
<tr>
<td>5B</td>
<td>Kukri_c1214_948</td>
<td>0.066</td>
<td>0.361</td>
<td>0.316</td>
<td>-0.044</td>
</tr>
<tr>
<td>6A</td>
<td>BS00065700_51</td>
<td>-0.050</td>
<td>0.399</td>
<td>0.454</td>
<td>0.056</td>
</tr>
<tr>
<td>6A</td>
<td>Kukri_c6128_238</td>
<td>-0.055</td>
<td>0.379</td>
<td>0.150</td>
<td>-0.229</td>
</tr>
<tr>
<td>7B</td>
<td>RAC875_c4876_2096</td>
<td>0.068</td>
<td>0.044</td>
<td>0.115</td>
<td>0.071</td>
</tr>
<tr>
<td>7B</td>
<td>wsnp_Ex_c8400_14157318</td>
<td>0.054</td>
<td>0.514</td>
<td>0.304</td>
<td>-0.210</td>
</tr>
</tbody>
</table>
Figure 5: The absolute difference between the allele frequencies between the CRI0 and CRI2 lines against their position on the chromosome.
5 Discussion

The aim of this study was to find out if genomic prediction could be used to predict crown rot resistance in a complex wheat population. Different models were used to predict CRI2 lines from the CRI0 lines and from a subset of the CRI2 lines. Prediction of the CRI2 lines based on the CRI0 lines was low, so appeared to be not useful in breeding new crown rot resistant varieties. Genomic prediction is possible within the CRI2 lines, yielding a high correlation between the predictions and the actual measured SB values. This can be used in future research to reduce the number of plants that need to be phenotyped, by using the genotype data. Two main explanations of the low predictive ability from the CRI0 to the CRI2 lines are explored below. One of the explanations is the change in allele frequencies between the two generations. The other explanations is the low correlation among different phenotype trials.

The heritabilities of the different phenotype trials were quite low and varied among the different trials. The correlation between the different trials was also very low. This indicates the genotypic variance was small compared to the unexplained variance, including the residual variance and genotype by experiment interaction. The genotype by experiment interaction might explain part of the variance and is an interesting topic in follow-up research.

Phenotyping crown rot infection and severity is very difficult. Some lines may be more resistant than others, but this effect is small in comparison to the noise. The trials from June 2016 and March 2016 had a high correlation and this might be due to the extra point that was added to the browning assessment scale. Adding an extra scale point allowed for death by crown rot to be identified as severe crown rot infection, instead of missing data with unknown cause. This also indicates the missing values in the phenotype trials might not be missing at random, but the data was partly missing because of severe crown rot. A valid question is whether the use of the extra scale point with the CRI0 lines would have resulted in more accurate predictions.

The GWAS was used for two different purposes: identifying important SNPs and exploring the effect of different kinship matrices. The important SNPs that were found and used for genomic prediction, based on the CRI0 lines were found on chromosomes: 1A, 1D, 2A, 3B, 4A, 5A, 5B, 6A and 7B, suggesting QTLs on these chromosomes. In the literature QTLs were identified on chromosomes 1A, 1D, 2B and 3B. The trait is complex and many loci on the genome are involved. Not all chromosomes with QTL from literature were found in this study. One of the causes of this could be the fact that the CRI0 lines already went through multiple selection steps and the places with the QTL might be not segregating anymore.

The different kinship matrices were constructed and compared in their performance in GWAS. This showed that building kinship matrices in a different way can have an effect on the results of the GWAS. The chromosome specific kinships yielded higher $-log(p)$ values and might have a better power of detecting
possible QTL’s. This is in line with earlier research (Rincent et al., 2014). In this case, the unmapped chromosome specific kinship matrices were used to do the GWAS, because that GWAS had the highest $-\log(p)$ values.

In cross-validation the predictive power within the CRI0 lines was explored. The CRI0 lines contained some predictive power, since a correlation of 0.29 was found between the predicted and measured values in the G-BLUP model. This low correlation shows the SNPs do not contribute much to the total variation. The CRI0 lines in the phenotype trials showed a very low correlation among each other, which could be a cause of the poor prediction.

In the prediction of the CRI2 lines the linear model with the ten selected SNPs from the GWAS, based on the CRI0 lines, did not have a high predictive power, since only 3% of the variance was explained. The correlation between the predicted and measured values was low (0.174). This model seems not powerful enough to help breeding crown rot resistant varieties. The prediction based on the CRI2 lines performed better, with a correlation of 0.271, but the model was less accurate in prediction than the G-BLUP and Gaussian kernel models.

Both the Gaussian and the G-BLUP model trained on the CRI0 lines did not perform well in predicting the CRI2 lines. The correlations were very low (0.050). This is lower than expected from the cross-validation. For breeders this is not useful in the selection of crown rot resistant lines. The CRI2 lines, on the other hand, can be used to predict the CRI2 lines. A sample of 200 CRI2 lines could predict the test set of CRI2 lines with a correlation of 0.409, which is quite accurate. This might be interesting for breeders. Instead of phenotyping all lines, a part of the lines can be grown and phenotyped and genomic prediction can be used to predict the outcome of the rest of the lines.

Prediction based on the combination of the CRI0 lines and the CRI2 lines did not perform much better than the model based on only the CRI2 lines, for the G-BLUP and the Gaussian kernel models. Only the linear model performed somewhat better (0.315), but the difference between the training sets was small. The CRI0 lines did not add much information to the prediction, which was expected from the poor prediction model based on the CRI0 lines. For G-BLUP and the Gaussian kernel model this could also be a result of the way the models work. In both techniques the phenotypic outcome is predicted based on relatedness between the lines. The CRI2 lines are more closely related and therefore determine most of the prediction.

The G-BLUP model and the Gaussian kernel model did not give very different results. The predictions from the Gaussian kernel were slightly more accurate, but no model was generally found to perform better in this study. This indicates the freedom of non-linear SNP effects did not increase the predictive power very much.

The changes in allele frequencies were explored to see whether they explain the inability to predict
the CRI2 lines from the CRI0 lines. Between the CRI0 lines and the CRI2 lines, two steps influenced the allele frequencies in the genome, the backcross and the selection steps. The backcross increases the allele frequency in the direction of the elite parental background. The elite parents are less crown rot resistant than the donor lines, so the backcross can have increased alleles that are not favorable in terms of crown rot resistance. After the backcross, the selection steps excluded the lines with low crown rot resistance and thereby increased the alleles associated with crown rot resistance. This way most alleles associated with crown rot resistance were preserved.

The allele frequencies of the selected SNPs were expected to increase for the SNPs with a negative effect on the SB score and are expected to decrease for SNPs with a positive effect, due to the selection steps. The results from the linear model showed this is not always the case. For one out of 14 SNPs, the allele change was as expected with random change (less than 0.025). Out of the remaining 13 SNPs, seven had an allele change as expected, in which the allele frequency was decreased for the SNP with a positive effect on stem browning and thus a positive effect on resistance (on chromosomes 1A, 1D, 5A, 5B, 6A and 7B). The six other SNPs changed in the opposite direction with an increase for the unfavorable allele (on chromosomes 3B, 4A, 5A, 6A and 7B).

This might be explained by the fact that the CRI0 lines were also selected on CR resistance. The backcross might have changed these important SNPs back from the selection that took place to create the CRI0 lines. The influence of these SNPs might have decreased by crossing multiple CRI0 lines to get to the CRI2 lines, so that the allele frequencies did not change much anymore after the backcross. Also, multiple donor lines were combined in the CRI2 lines, which could have caused interactions between alleles that make an allele less favorable.

Allele frequency changes were found everywhere on the genome. This is partly because of the backcross, but it could also be an indication that crown rot resistance is a trait influenced by multiple parts of the genome. It would be interesting to explore which alleles were increased due to the backcross, and which were increased or decreased due to the selection steps. This could be achieved by adding pedigree information to the model. It could give an interesting insight in possible QTL positions.

The change in allele frequencies could be one of the reasons for the poor quality of the genomic prediction of the CRI2 lines from the CRI0 lines. The alleles that have an impact on crown rot resistance in the CRI0 lines, might have decreased in the CRI2 lines, resulting in a much smaller effect on the CRI2 lines. The selection steps can push some alleles out of the population, in which case there is no variation based on these alleles anymore. So an allele that greatly impacts the CRI0 lines, might have almost no impact anymore on the CRI2 lines. This way, if the allele frequency of an allele comes close to zero, the genetic effect of this allele also disappears from the population. The QTL found in CRI0, might have no or little effect in the CRI2 lines. The trait is influenced by multiple smaller QTLs, so the changes over
the whole genome could have impacted the effects of the SNPs.

The models used in this study can be used to predict crown rot resistance within a population. The CRI2 lines cannot be accurately predicted from the CRI0 lines using the techniques explored in this study. The unexplained variance in the phenotyping is quite large and the effects of the SNPs might have changed because of the large differences in the allele frequencies between the two generations. The prediction might have been more accurate if the scale point 5 was added to the stem browning scale for the CRI0 trials. The CRI2 lines can be accurately predicted on a subset of the CRI2 lines, using a G-BLUP or Gaussian kernel model. This could help in reducing the number of lines that need to be grown and phenotyped.
References


