Prediction of cross performance in barley

Promotor: dr.ir. P.Stam
Hoogleraar in de plantenveredeling
in het bijzonder de selectiemethoden
en resistentieveredeling

Co-promotor: dr.ir. I. Bos
Wetenschappelijk hoofdmedewerker
Departement plantenveredeling en gewasbescherming

Johan.W. Schut

Prediction of cross performance in barley

Proefschrift

ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van de Landbouwuniversiteit Wageningen,
dr. C.M. Karssen,
in het openbaar te verdedigen
op donderdag 10 November 1998
des namiddags te vier uur in de Aula

The research described in this thesis was supported by the Wageningen Agricultural University (WAU), The Netherlands. The research was carried out at the Department of Plant Breeding (WAU) in cooperation with Cebeco Zaden BV, Lelystad, Vanderhave Research, Rilland, Zelder BV, Ottersum and CIMMYT, Mexico

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG.

Schut J.W.

Prediction of cross performance in barley / J.W. Schut -[S.I.: s.n.]
Thesis Wageningen Agricultural University -with references- with summaries in English and Dutch.
WAU-Laboratory of Plant Breeding, P.O. Box 386, NL-6700 AJ Wageningen, The Netherlands
ISBN 90-5485-936-9

Bibliographic abstract:

In this thesis opportunities for prediction of cross performance in a plant breeding program are investigated. For this research 20 SSD-line populations from crosses between European two-row spring barley lines were evaluated for four quantitative agronomic traits in seven environments, divided over two years. The midparent value appeared to be a good predictor of average offspring performance and useful in practical breeding. However, for most crosses the midparent value for grain yield overestimated the offspring average. The relatedness between parents was expected to predict the variance among the offspring. However, predictions using genetic distances based on pedigree data, morphological trait data, and AFLP-marker data, performed poorly. A genetic distance based on AFLP-markers associated with the trait variation among the parents gave a somewhat better prediction. The correlation between the parental responses to different environments appeared to be a reasonable predictor of grain yield among the offspring. Other variance predictors based on parental differences for agronomic trait data or early generation (F4) variance among the offspring, mainly predicted variance resulting from segregating major genes. These genes are often fixed in practical breeding programmes and therefore not very relevant. Grain yield data from small three-row plots in an early generation evaluation did not correspond with large plot yield data due to interplot competition.

Keywords: additive main effects and multiplicative interaction, coefficient of coancestry, coefficient of parentage, cross prediction, early generation selection, genetic map, genetic similarity, genotype-by-environment interaction, *Hordeum vulgare*, marker selection, progeny variance, segregation analysis, stability

Contents

1

General introduction

A plant breeding programme generally consists of the creation of novel genetic variation, the subsequent selection of new cultivars and the propagation and maintenance of these cultivars. In many breeding programmes the genetic variation is created by crossing genetically divergent parents. The choice of these parent combinations is very important. It decides on which part of the initially available genetic variation new cultivars will be based and which genes will be (re)combined by crossing. It can be regarded as the first selection step. The resulting genetic variation is the determining factor for the offspring performance, which is defined as the level of compliance of the offspring with preset breeding goals.

In many cases the offspring performance is assessed directly by the performance of a hybrid (e.g. in maize, cabbage or tomato), a population (e.g. in rye), or a clone. For clones performance as a cultivar (e.g. in potato or rose) can be distinguished from performance as a cultivar parent (e.g. in ryegrass). In some cases offspring performance is assessed indirectly, after several generations of inbreeding, in a segregating population. Then the population performance is evaluated by the probability of selecting a recombinant inbred line that performs well, either as a cultivar (e.g. in wheat, barley or lettuce) or as a hybrid parent. The recurrent nature of the breeding process is demonstrated by the fact that an important element of the offspring performance is its performance as a parent in the next breeding cycle. The effects of the environment on the offspring performance can be quite large. The relevant influences for plant breeders are captured in the concept of genotype by environment interaction. It describes the change in the differences between genotypes when going from one environment to the other. Breeders cope with this interaction by breeding different cultivars for different environments or by breeding cultivars that combine good performance with high stability over the different environments.

The choice of parent combinations in a breeding programme is based on the breeder's knowledge of the performance of the individual parents including their performance as a parent in earlier breeding cycles. Many traits are considered simultaneously, of which some are based on single genes, while others have a complex polygenic basis. Usually parent combinations are chosen in such a way that weaknesses of one parent are compensated for by the other parent and vice versa. A second consideration is the degree of heterosis expressed by a hybrid, or the degree of transgression expressed by a segregating offspring population. The decision of the breeder on which parent combinations will be chosen depends on his expectation about the offspring performance. In view of the large resources allocated to making crosses and evaluating offspring

performance, a good cross prediction is of prime importance. The final choice of parent combinations is often made on the basis of implicit expert knowledge, often referred to as the 'breeder's eye'. However, the decision is usually supported by predictions of offspring performance based on explicit information. These cross predictions can be classified by the type of information that is used:

A. predictions based on known genes for the relevant traits, including the parental genotypes, i.e., the allele constitution of the potential parents.

B. predictions based on information about the candidate parents that can be obtained before using them as a crossing parent, e.g. geographic origin, pedigree, and trait data. These traits may range from agronomical and morphological traits to biochemical (e.g. isozymes and storage proteins) and molecular traits (e.g. DNA-based markers, like RFLP, RAPD and AFLP).

C. predictions based on 'past parent performance' of the candidate parent, often obtained by testcrosses

D. predictions based on a relatively inexpensive assessment of a limited offspring population of the candidate cross. From this experiment population parameters like mean and variance can be predicted. These predictions are only applied in the case of subsequent selection within the offspring population, e.g. in selffertilising crops like barley, wheat and lettuce and in crossfertilising, clonally propagated crops like potato and strawberry.

The four classes are briefly discussed in the following paragraphs. If necessary, we take into account a subclassification that can be made on the basis of the predicted breeding behaviour, e.g. mean offspring performance, heterosis, or genetic variance among the offspring.

For most monogenic traits, e.g. many disease resistances, and some oligogenic traits, e.g. flower colour in several ornamental species, the underlying genes are known. For other traits, which are mainly polygenic and quantitative, i.e., measured at an ordinal or a continuous scale, a QTL-analysis (Quantitative Trait Loci: Lander and Botstein, 1989; Jansen, 1992) can shed some light on the genetics underlying the traits and the parental genotypes. Van Berloo and Stam (1998) present an example of a QTL-based cross prediction for flowering time in *Arabidopsis thaliana*. However, usually only part of the genetic variation can be explained by QTL due to simplifying assumptions in the QTL-model and noise in the data. Besides, the prediction of offspring performance on the basis of genetic markers linked to QTL can be seriously hampered by epistasis, also known as 'genetic background effects', and lack of linkage disequilibrium between markers and QTL across a set of potential parents. Therefore, the results of a QTL analysis based on one cross cannot always be extrapolated to other crosses. These are probably

the main reasons why in many cases a QTL-based prediction of offspring performance for complex traits is still insufficiently accurate and/or too expensive.

In the case that the genetics of the traits and the parental genotypes are largely unknown other methods of cross prediction can support the breeder's choice of parent combinations. They are classified above as B, C and D. These methods of cross prediction mainly arise from quantitative genetic theory. They are largely based on the relationship between parents and offspring, either inferred empirically or based on genetic theory. These types of cross prediction are usually performed for one or more quantitative polygenic traits. Examples of such traits are: yield, partial resistance, concentrations of desired and undesired substances in the harvested product (protein, oil, sugar, starch, nitrate). But also combining abilities (established by test crosses), indices (e.g. financial yield) and other derived statistics (e.g. stability parameters (Jinks and Pooni, 1980; Lin and Binns, 1991)) can be considered as quantitative traits.

The mean offspring performance can be predicted before making crosses with the candidate parents (B). This prediction is often based on the midparent performance, i.e., the average performance of the parents (Bos and Caligari, 1995). A typical example is the situation in which a breeder is trying to combine parents with complementary strengths and weaknesses. The aim is a descendant without undesired levels of performance for any of its traits. In cases where midparent performance predicts poorly, components or associated characters may be observed on the parents and successfully applied in mean prediction. An example is presented by Neele (1990) for tuber yield in potato. Bos & Sparnaaij (1993) present a prediction method based on analysis of trait components.

Prediction methods based on parent data before making crosses with the candidate parents (B) may also concern heterosis and genetic variance among the offspring. Both these types of breeding behaviour are related to F1 heterozygosity. Heterosis is a direct result of number of heterozygous loci in the F1 and their dominance effects. Genetic variance in an inbred offspring population is an indirect result of the number of heterozygous loci in the F1 and their additive effects. The number of heterozygous loci in the F1 is assumed to be associated with the relatedness of the parents. Therefore, many authors have used parental relationship measures to predict heterosis or genetic variance (brief overviews by Cowen and Frey, 1987b; Loiselle et al., 1991; Stuber, 1992; Charcosset and Essioux, 1994). These parental relationship measures can be established using three sources of information from the parents: geographic origin, pedigree and trait data, as mentioned above. The expected association between relatedness and heterosis or genetic variance is based on the assumption that the average effect of each heterozygous locus in the F1 is more or less equal for the different parent combinations. Deviations of this assumption, as well as other sources of error, make results quite variable. Several studies on genetic variance prediction in selffertilising crops report a relatively inaccurate estimation of this aspect of breeding behaviour. This may be caused by a small number of crosses, a small number

of lines per cross, a small-scale evaluation in a rather early generation (e.g. F3-F4), or a combination of these factors (Moser and Lee, 1984; Manjarrez-Sandoval et al., 1997; Burkhamer et al., 1998).

The prediction of offspring performance may be improved by information about the 'past parent performance' of the candidate parent (C). The general combining ability (gca) of a parent can be determined by the evaluation of a set of test crosses with that parent. This approach is often used in breeding programmes for maize hybrids and for potato or strawberry clones. An example of this method is presented by Neele et al. (1991) for potato tuber yield.

For heterosis prediction Bernardo (1994) proposes to use a limited sample of predictor hybrids descending from the candidate parents. He also proposes the use of parental relationship measures to estimate the genetic covariance matrix between the observed hybrids and the ones to be predicted. Combining the predictor hybrid data with the estimated covariance matrix, he derives best linear unbiased predictors (BLUPs) using a mixed model. Instead of a mixed model Charcosset et al. (1998) use a factorial regression model for which the parental relationship matrix is transformed into a limited set of regressor variables. Both approaches appear to predict well, especially in the case of unrelated parents, when tested on maize forage yield data (Charcosset et al., 1998). Both procedures also involve a gca-component, thus combining the gca and heterosis prediction.

If selection within an offspring population is necessary to obtain a new cultivar or cultivar parent, a breeder can also directly assess the offspring of a potential parent combination (D). A limited offspring population in an early generation and/or a relatively inexpensive assessment method may provide an indication which parent combinations should be chosen to proceed to a more extensive crossing and/or selection programme. Examples of this procedure for the prediction of offspring mean and variance are presented by Jinks and Pooni (1980) using F3-lines of tobacco and by Caligari and Brown (1986) using second year clones of potato. Progeny variance can also be predicted on the basis of heterosis in the F1. To combine predicted mean and variance of a segregating population Jinks and Pooni (1976) propose to estimate the proportion of offspring genotypes that would exceed an arbitrary threshold, assuming a normal distribution of trait values. Crosses are selected on the basis of this parameter. This combined measure of performance of a cross was used by Van Ooijen (1989b) in a study to assess its predictive value when mean and variance estimates are based on small plots in early generations of spring wheat. Another combined selection parameter may be the expected value of the best performing offspring genotype in a population of a certain size, again assuming a normal distribution.

Many approaches to predict the offspring performance of a certain parent combination can be applied for different traits in different crops. Ideas behind these approaches can often be extended to crops with different modes of reproduction.

Objectives and outline of present study

The subject of the present study is the choice of parent combinations in an inbred crop using predictions of offspring mean and variance. The first objective is the comparison of several existing methods of mean and variance prediction. A second objective is the investigation of their usefulness for practical breeding. A third objective is the investigation of several modifications that are proposed to improve the usefulness of variance prediction. Among them is the use of a relatively new source of parental relationship information: AFLP-markers (Vos et al., 1995).

For this study several agronomic traits, like grain yield and plant height, are observed in parents and offspring populations of European two-row spring barley (*Hordeum vulgare* L.). Barley is chosen as a model crop, because of several considerations: 1) it is diploid and self-fertilising, which simplifies some of the assumptions that have to be made in the genetic models; 2) it has a short generation length, so several generations per year could be raised, if necessary; 3) it has often been used in applied genetical studies, so much genetic information is available; 4) it is an important agricultural crop, with the fifth largest cultivated area in the world. The parent lines were chosen to represent the population of parents employed in commercial barley breeding programs in Northwest Europe over the last 20 years. They are rather closely related, primarily as a result of breeding for malting quality and regional adaptation. Genetically distant material is mainly used to introduce disease resistances by backcrossing procedures and it is not expected to have significantly contributed to other traits.

In several previous studies the number of crosses and the number of environments used do not allow general conclusions to be drawn with respect to the predictability of cross performance (Cowen and Frey, 1987b; Moser and Lee, 1984; Helms et al., 1997; Manjarrez-Sandoval et al., 1997; Burkhamer et al., 1998). Further it is mentioned that inaccuracy of mean and variance estimates hampers the drawing of clear conclusions. In order to remove these drawbacks in the present study we aimed to control, within the limits of experimental feasibility, several factors that influence the reliability of the correlation between predicted and observed cross performance. First, we use a relatively high number crosses. These 20 crosses are derived from 18 different parent combinations plus two randomly chosen reciprocals. The parent combinations are based on a partial diallel crossing design using 18 parents (n=18; s=2; Kempthorne and Curnow, 1961). Second, each cross is represented by 48 recombinant inbred lines (RILs) produced by single seed descent (SSD). This enables reliable between RIL variance estimates, provided the individual RIL performance is estimated accurately. Third, single seed descent is extended until the F5 generation so as to achieve a high level of homogeneity within the lines. Fourth, large plots, similar to the ones in commercial breeding programs, are used. In combination with incomplete block designs these decrease the error variance. Fifth, for the sake

of generalisation and in order to investigate genotype by environment interaction, the RILs have been evaluated in seven environments, distributed over two years. Thanks to the kind support of three Dutch breeding companies (Cebeco, Lelystad; VanderHave, Rilland; Zelder, Ottersum) we could add three locations to the two university sites in Wageningen (Unifarm) and Swifterbant (Ir. A.P. Minderhoudhoeve).

In chapter 2 AFLP-markers (Vos et al., 1995) are investigated as a source of parental relationship information. AFLP-based genetic similarity estimates are compared with parental relationship measures based on pedigree and morphological trait data. A bootstrap procedure is presented that approximates the inaccuracy of the correlation coefficient between two relationship measures. Further, we discuss the usefulness of AFLP-based genetic similarities for cultivar identification and for assessment of genetic diversity.

In chapter 4 the AFLP-based genetic distances are tested for their usefulness in variance prediction. Their predictive value is compared with that of parental relationship measures based on pedigree, agronomic and morphological trait data. Combinations of these relationship measures are also examined for their association with progeny variance. We investigate the effect of 'major genes' on the variance predictions.

The unconditional use of all AFLP-markers in genetic distance estimation may cause a lack of representation of the relevant genes for a trait. Map information can be used to weight markers for marker density in a genetic distance calculation. This is expected to remove overrepresentation of chromosome regions with a high marker density. Another approach is the use of only those markers that show a strong association with a trait in the parent population. Variance prediction based on the resulting genetic distance estimates is examined in chapter 5.

Progeny variance and mean for yield are also predicted on the basis of early generation (F4) offspring evaluation in small plots. This is described in chapter 3. The prediction of mean RIL performance by midparent values is also examined. Effects of 'major genes' and interplot competition are discussed, as well as the influence of genotype by environment interaction.

Genotype by environment interaction for yield is further investigated in chapter 6. Several stability parameters are calculated for parents and offspring. The usefulness of midparent values to predict mean RIL stability statistics is discussed. A biplot representing part of the nonadditivity is used to demonstrate the relationship between parents and offspring. Further, we use the correlation between the parental environment-specific response vectors (Habgood, 1977) to predict progeny variance for yield. The environment-specific responses are the residuals from an analysis of variance using a model with additive genotype and environment effects.

Finally, in chapter 7 the main results are discussed with regard to applications in practical plant breeding and with regard to topics for further research.

2

Association between relationship measures based on AFLP-markers, pedigree data and morphological traits in barley [1]

Johan W. Schut, XiaoQuan Qi, Piet Stam

Abstract

Thirty one barley lines were used to investigate the agreement of three relationship measures: genetic similarities based on 681 AFLP-markers, coefficients of coancestry based on pedigree data and generalised distance based on 25 morphological characters (morphological distance). Bootstrap analysis was used to estimate the accuracy of the correlation estimates. AFLP-based genetic similarities showed a poor-to-moderate correlation with coefficients of coancestry within the core set of twenty five European two-row spring barleys. Morphological distance was not significantly correlated with either genetic similarity or coefficient of coancestry. The precision of all correlation coefficient estimates, however, was low. Inclusion of two European winter barleys, two North American two-row spring barleys, and two North American six-row spring barleys in the AFLP-analysis resulted in a much stronger correlation between genetic similarity and coefficient of coancestry. This suggests good opportunities for the use of AFLP-markers to assess genetic diversity by distinguishing between the major ecotypes of barley. Besides, each of the eight primer combinations used in the AFLP-analysis was able to identify all 31 lines uniquely, showing the usefulness of AFLPs for cultivar identification. Because of the inaccuracy of the investigated relationship measures, resulting in low values of the correlation coefficient estimates, prediction of the breeding behaviour of parent combinations may be improved by the use of a combination of relationship measures, thus decreasing the effect of their individual independent errors.

Key words: bootstrap analysis, coefficient of coancestry, cultivar identification, genetic similarity, *Hordeum vulgare*

Introduction

[1]published in: Theor Appl Genet (1997) 95:1161-1168

Knowledge about relationships between genotypes that may be used in new crosses and about genetic diversity in available germplasm is very useful for plant breeders. It supports their decisions on the selection of crossing combinations from large sets of parent genotypes and is helpful when they want to widen the genetic basis of a breeding program. The selection of crossing combinations is supported by prediction of the performance of offspring resulting from crossing combinations between inbred parents. Cowen & Frey (1987b) distinguished three types of breeding behaviour that can be predicted: heterosis, transgressive segregation and genetic variance among offspring. However, higher transgressive segregation fractions are a direct result of higher genetic variances among offspring, taking into account the difference between the performances of the parents. Therefore, they do not have to be considered as separate phenomena. On the other hand, heterosis and genetic variance are a direct and an indirect result of heterozygous loci in the F1 and their effects. The degree of relationship between two parent genotypes is mainly expected to predict the number of heterozygous loci in their hybrid.

Knowledge about genotype relationships is usually based on three sources of information: (1) geographic information about the origin of genotypes, (2) pedigree information and (3) information about plant characteristics. Geographic information is helpful in most cases. It is specifically used when other information on genotypes is not available or very sparse. This is often the case for gene-bank material. Pedigrees of varieties and breeding lines are often well documented. They trace back to landraces and wild accessions. However, pedigrees sometimes contain erroneous or incomplete information. Plant characteristics are the only source of relationship information that is, or can be made available, for any set of genotypes. Such characteristics can be divided into four arbitrary groups: agronomic characters, morphological characters (used to distinguish between varieties), biochemical characters (e.g. storage proteins, isozymes) and molecular (DNA) markers. Differences between genotypes with regard to any of these characteristics are either indirect or direct representations of differences at the DNA level and are therefore expected to provide information about genetic relationships.

A range of measures is available to quantify relationship information. For pedigree information Malécot (1948) presented the coefficient of coancestry (f), also known as kinship coefficient or coefficient of parentage. For agronomic and morphological traits measured at a continuous or ordinal scale one can use multivariate statistical techniques and construct a p-dimensional space, where p is the number of traits. The Euclidean distance between the points representing the genotypes may be used as a measure of relatedness (Goodman, 1972). Generalised distance is an extension of Euclidean distance correcting for correlation between traits (Mahalanobis, 1936). Plant characteristics, like isozymes and molecular markers, are scored as binary data. A commonly used similarity measure was presented by Dice (1945). Nei and Li

(1979) demonstrated the usefulness of this genetic similarity for isozyme and molecular-marker data.

In winter wheat Cox et al. (1985b) found a poor correlation between coefficient of coancestry and genetic distance based on storage protein data. Genetic distance based on combined isozyme and morphological data showed a moderate correlation with pedigree based distance in soybean (Cox et al., 1985a). However, Souza and Sorrells (1991a,b) concluded that distance measures based on quantitative and qualitative morphological characters (the latter including isozyme characters) in oats did not correspond very well with pedigree data. The introduction of molecular markers like RFLP (restriction fragment length polymorphism; Botstein et al., 1980) and RAPD (random amplified polymorphic DNA; Williams et al., 1990) created the opportunity to assess genetic relationships directly at the DNA level. A priori, similarities at the DNA level are expected to be in better agreement with pedigree information than similarities based on morphological traits or gene products, whose expression can be influenced by the environment and/or epistatic interactions. However, results based on RFLP- and RAPD-markers are quite variable in this respect. Tinker et al. (1993) showed a moderate correlation between RAPD-based genetic distance and coefficient of coancestry when considering 27 Canadian spring barley lines. Graner et al. (1994) used RFLP data to estimate genetic distances between 48 European barley varieties. They found poor-to-moderate correlations between marker-based distance and coefficient of coancestry. Correlations were higher in spring barley than in winter barley. Autrique et al. (1996) obtained a moderate correlation between RFLP-based genetic distance and distance based on sixteen agronomic and morphological traits in durum wheat. Correlations between these two relationship measures and coefficient of coancestry were poor.

Recently AFLP, a PCR-based molecular marker technology was introduced by Vos et al. (1995). Using PCR-amplification genomic restriction fragments are selectively multiplied to adequate detection levels producing reproducible DNA-fingerprint patterns. The fast and reliable production of many marker data points is an advantage of AFLP over RFLP and RAPD.

The aim of the present study is to investigate the agreement of AFLP based genetic similarities, coefficients of coancestry and generalised distance based on morphological characters. Special attention will be paid to the precision of the correlation estimates. We will discuss, in brief, the usefulness of AFLPs for variety identification as well as opportunities for the prediction of breeding behaviour of crosses and the assessment of genetic diversity.

Materials and methods

Plant materials

Thirty one barley (*Hordeum vulgare* L.) lines were used in this study. The core set consisted of 25 European two-row spring barley varieties and breeding lines. They were chosen to represent parent populations employed in commercial spring barley breeding programs in Northwest Europe over the last 20 years. Firstly, relationship measures were compared for this core set. Secondly, a set of six cultivars consisting of two European winter barleys (one two-row; one six-row) and four North American spring barleys (two two-row; two six-row) were added as representatives of some other major barley groups. This offered an opportunity to investigate possibilities for the assessment of genetic diversity using AFLPs. Names and details of the 31 lines are presented in Table 2.1.

AFLP analysis

DNA extraction followed the CTAB-method described by Van der Beek et al. (1993).

The AFLP-technique is described by Vos et al. (1995): DNA-restriction uses the enzyme combination *Eco*RI/*Mse*I. After adapter ligation DNA-fragments are amplified using PCR. Primer annealing is targeted at the adapter and restriction site sequence. Three-nucleotide extensions on both *Eco*RI and *Mse*I primers cause selective amplification of fragments. The AFLP-analysis followed the protocol described by Van Eck et al. (1995) with modifications by Qi and Lindhout (1997).

Primer combinations were chosen that produce a high number of unambiguous polymorphisms in a wide range of barley germplasms (Qi and Lindhout, 1997). The eight primer combinations that were used are presented in Table 2.2.

Genetic-similarity estimation

AFLP-bands were scored as present (1), absent (0) or as a missing observation (–1) for the different genotypes. Often several AFLP-markers within a primer combination show pleiotropic behaviour or very close linkage (Qi and Lindhout, 1997). Likewise, in our set of genotypes polymorphic markers with identical polymorphism patterns were found within primer combinations. We also found markers within primer combinations that seemed to be allelic. In all of these cases a second marker does not add any new independent information to a genetic-similarity estimate. Therefore these redundant polymorphic markers within primer combinations were discarded before calculating genetic similarities.

Table 2.1. Genotypes used in AFLP analysis, their pedigree, country of origin, type (2=two-row; 6=six-row; s=spring barley; w=winter barley), the possibility to trace the pedigree to original ancestors (x-mark means: more than 75% of the pedigree can be traced to original ancestors) and the availability of morphological trait data (x-mark means: data available). The dashed line divides the core set of European two-row spring barleys from the rest

Genotype	Pedigree	Country of origin	2-row/ 6-row type	spring/ winter type	>75% known pedigree	morphological trait data available
Apex	Aramir x (CEB-6711 x (Julia(3) x (Volla x L-100))	the Netherlands	2	s		x
Aramir	Volla x Emir	the Netherlands	2	s	x	
Baronesse	5238/8-74 x 754465	Germany	2	s		x
Bonaire	E-77040-8107 x (CEB-8188 x Apex)	the Netherlands	2	s	x	x
CEB-9079	Robin x (CEB-8498 x Efron)	the Netherlands	2	s	x	x
CEB-9186	CEB-8187 x Golf	the Netherlands	2	s	x	x
Drossel	(FLO-1625/56 x Union) x Ingrid	Germany	2	s	x	x
Forester	CSBM2 x Sherpa	UK	2	s		x
GEI-119	Aramir-EI x Goldmarker(3)	the Netherlands	2	s		x
Georgie	Vada x Zephyr	UK	2	s	x	x
Gunhild	(Algerian x Lone) x MGH-63199	Denmark	2	s	x	x
IVP9211327	(GEI-119 x Gunhild) x (Prisma x Apex)	the Netherlands	2	s	x	
IVP9211510	(Prisma x Apex) x (GEI-119 x Gunhild)	the Netherlands	2	s	x	
Karat	K-1443-70 x I-2931170	Czech Republic	2	s	x	x
Kenia	Binder x Gull	Denmark	2	s	x	
Midas	((Proctor x Wong) x mildew-res.A) x mutant of Maythorpe	UK	2	s	x	x
Nudinka	Emir x Weihenstephan 1606 Nackt	Germany	2	s		
Porthos	Lignee-207 x Emir	France	2	s		x
Prisma	(Trumpf x Cambrinus) x Piccolo	the Netherlands	2	s	x	x
Proctor	Kenia x Plumage Archer	UK	2	s	x	
Riff	(VDH-240-79 x Karat) x Apex	the Netherlands	2	s	x	x
Triangel	((Mazurka x Ofir) x SVP-6045-66/25) x ((Villa x (Agio x Piroline)) x Carlsberg)	the Netherlands	2	s	x	x
Vada	Hord.laevigatum x Gull	the Netherlands	2	s	x	
Yriba	(Maris Yak x (Rika x Baladi-16-133)) x Rika	France	2	s	x	x
ZE-87-3414	Efron x (Aramir-EJ x Iraq-10922)	the Netherlands	2	s	x	x
Franka	((Vogelsanger Gold x Senta) x (Dura x Dea)) x Vogelsanger Gold	Germany	6	w	x	
Harrington	Klages x ((Gazelle x Betzes) x Centennial)	Canada	2	s	x	
Igri	(Malta x Carlsberg 1427) x Ingrid	Germany	2	w		
Morex	Cree x Bonanza	USA	6	s	x	
Steptoe	Wash.Sel.3564 x Unitan	USA	6	s		
TR-306	(Abee x TR451) x WM 793-1776	Canada	2	s	x	

Table 2.2. Primer combinations used in AFLP analysis

Number	E+3/M+3 nucleotide extensions
E33M54	E+AAG/M+CCT
E33M61	E+AAG/M+CTG
E35M54	E+ACA/M+CCT
E35M61	E+ACA/M+CTG
E38M50	E+ACT/M+CAT
E38M59	E+ACT/M+CTA
E38M60	E+ACT/M+CTC
E38M62	E+ACT/M+CTT

The genetic similarities (*gs*) are calculated following Nei and Li (1979) :

$$gs_{ij} = \frac{2N_{ij}}{N_i + N_j}$$

where N_{ij} is the number of bands present in both genotypes *i* and *j*, N_i is the number of bands present in genotype *i* and N_j is the number of bands present in genotype *j*. In the case of a missing observation for a marker in genotype *i* and/or *j*, this marker was not included in the calculation of gs_{ij}. The accuracy of *gs*-estimates as influenced by sampling and missing marker data was assessed by taking bootstrap samples (Efron & Tibshirani, 1993) from all 681 markers, including polymorphic as well as monomorphic markers. Bootstrap standard-deviation estimates were based on 1000 samples.

Principal-coordinate analysis (Gower, 1966) was used to obtain a graphic representation of the relationship structure of the thirty one genotypes. Computations were performed using the MDS-procedure in SAS (SAS Institute Inc., 1992).

Pedigree analysis

Pedigrees of the genotypes were gathered from several sources in literature (Baum et al., 1985; Arias et al., 1983) and from personal communication with breeders and researchers. The coefficient of coancestry *f* between two genotypes, as defined by Malécot (1948), was calculated. This is the probability that a random allele at a random locus in one genotype is identical by descent to a random allele at the same locus in the other genotype (Cox et al, 1985b). A FORTRAN-program obtained from Van Hintum (CGN, CPRO-DLO, Wageningen) was used to calculate *f*. The underlying assumptions are given by Van Hintum and Haalman (1994): (1) a genotype receives half its genes from each parent; (2) parents involved in crosses are homozygous and homogeneous; (3) ancestors for which no pedigree is available are

unrelated; (4) if selections are made from a cultivar, this cultivar is assumed to be the variable offspring of a cross between two unrelated lines. A selection from the cultivar is one of the offspring lines; if the cultivar itself is said to be used as a parent in a cross, then in fact one of the offspring lines has been used.

Genotypes often lacked some pedigree information. For only 23 genotypes (Table 2.1), of which 19 were core set genotypes, more than 75 % of the pedigree could be traced back to original ancestors, e.g. landraces. The f-value of a combination of 2 of these 23 genotypes was defined as 'well known' (f_{wk}) or complete (Graner et al., 1994). Also the f-values of two combinations between a parent and its offspring line were defined as 'well known', despite the fact that 75% or less of the parent pedigree could be traced back to original ancestors. Due to the direct relationship in this type of combination, the lack of pedigree information about the parent genotype does not have a strong effect on the f-value.

Morphological trait analysis

Out of 34 morphological traits in barley described by the international union for the protection of new varieties of plants (UPOV, 1981), we had at our disposal data on 25 traits in only 18 lines (Table 2.1) from the core set of 25 European two-row spring barleys. These data were obtained at our Wageningen site in 1994 in the presence of the relevant UPOV reference cultivars using one-row plots and two replicates. The data were confirmed by a similar trial in 1996. The traits are listed in Table 2.3.

The observed data were standardised per trait and a principal components analysis was performed. The principal components having an eigenvalue greater than an arbitrary value K=1.0, were used to calculate the generalised distances (morphological distance, *md*) between the lines (Goodman, 1972).

Bootstrap analysis of correlation coefficients

Simple (r) and rank (r_s) correlation coefficients between genetic similarities (*gs*), coefficients of coancestry (f) and morphological distances (*md*) were calculated. To test whether correlations were significant we used a bootstrap procedure (Efron and Tibshirani, 1993) to estimate 95%-confidence intervals for r. Bootstrap samples were produced by sampling with replacement from the set of genotypes (Schut, 1997). Then the *gs*-, f-, and *md*-matrices were constructed with rows and columns based on the genotype bootstrap sample. Due to resampling of the same genotype, some matrix cells contained a similarity or a distance between a genotype and itself. The contents of these cells were discarded before the calculation of the bootstrap correlation coefficient. For each correlation coefficient a 95%-confidence interval was constructed based

on 2000 bootstrap samples. The BC_a method (Efron and Tibshirani, 1993) was used to correct for bias and unequal variance to obtain a higher accuracy of the interval estimation.

Table 2.3. Twenty five morphological traits (UPOV, 1981) used to calculate morphological distances

Number	Trait
1	Plant: growth habit
2	Lower leaves: hairiness of leaf sheaths
3	Flag leaf: attitude
4	Flag leaf: anthocyanin colouration of auricles
5	Flag leaf: intensity of anthocyanin colouration of auricles
6	Flag leaf: glaucosity of leaf sheath
7	Time of ear emergence (first spikelet visible in 50% of ears)
8	Awns: anthocyanin colouration of the tips
9	Awns: intensity of anthocyanin colouration of the tips
10	Ear: glaucosity
11	Ear: attitude
14	Ear: shape
15	Ear: density
16	Awn: length compared to ear
18	Rachis: length of first segment
19	Rachis: curvature of first segment
20	Rachis: humping of segments (in mid-third of ear)
22	Sterile spikelet: attitude
23	Sterile spikelet: length of lemma
24	Sterile spikelet: shape of tip
25	Median spikelet: length of glume and awn relative to grain
28	Grain: anthocyanin colouration of nerves of lemma
29	Grain: spiculation of inner lateral nerves of lemma
30	Grain: hairiness of ventral furrow
31	Grain: disposition of lodicules

Results

Genetic-similarity estimation

In total 681 markers were used to estimate genetic similarities and 43.3 % of them showed polymorphism in the complete set of 31 genotypes. Restricting the set to 25 European two-row spring barleys yielded a smaller percentage of polymorphic markers: 37.9 %. However, each of the eight primer combination sets of markers could discriminate all thirty one barley genotypes.

Genetic similarities among all genotypes ranged from 0.857 to 0.978 with mean 0.919. Within the group of European two-row spring barleys the average *gs* was 0.932 ranging from 0.901 to 0.978.

Principal-coordinate analysis resulted in a three-dimensional graphic representation of the relationships between the genotypes (Figure 2.1). The correlation coefficient between genetic similarities and Euclidean distances in the graph was −0.86.

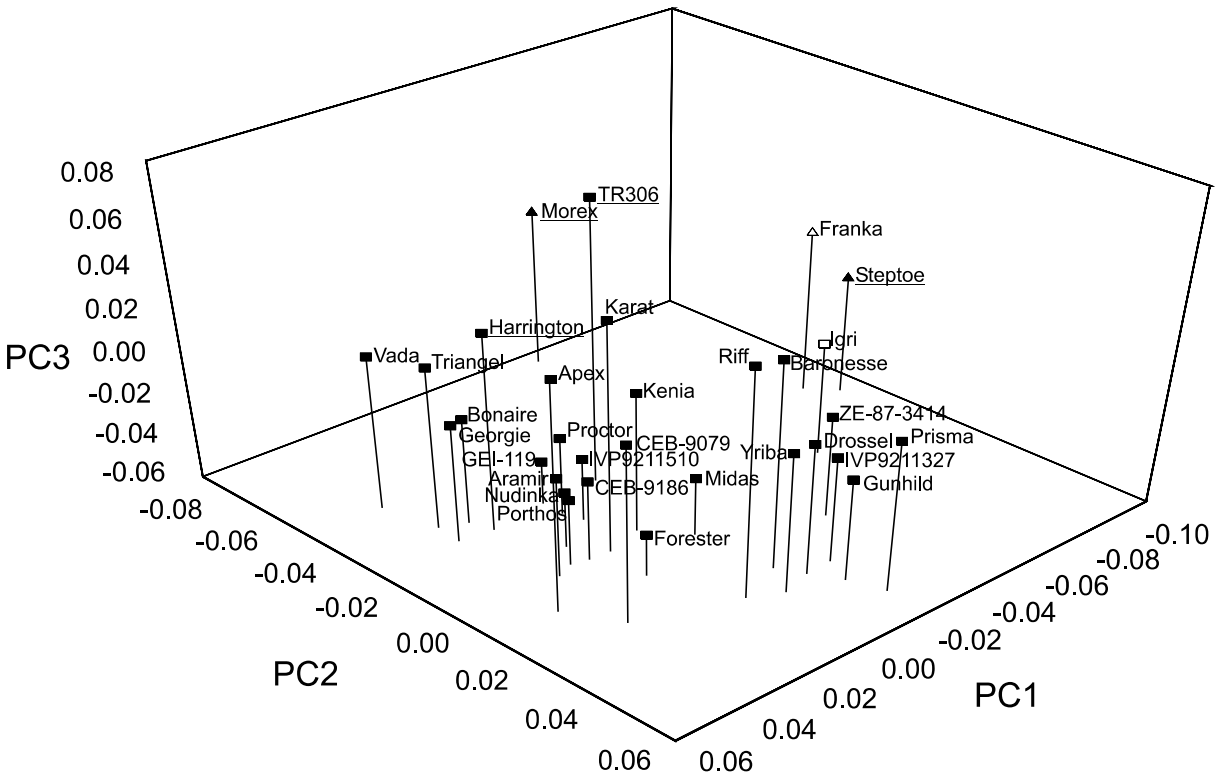

Figure 2.1. Relationships between 31 barley lines visualised by principal-coordinate analysis using AFLP based genetic similarities. ■=two-row spring type; ▲=six-row spring type; □=two-row winter type; △=six-row winter type. North American lines are *underlined*. *PC1*, *PC2* and *PC3*: first, second and third principal coordinates.

Pedigree analysis

Coefficients of coancestry that were defined as 'well known' (f_{wk}), ranged from 0 to 0.623 with mean 0.132. Within the core set of European two-row spring barleys, f_{wk} had an average of 0.176 ranging from 0.039 to 0.623.

Morphological trait analysis

After standardisation and principal component analysis, the first ten principal components, explaining about 87% of the variation, were used to calculate morphological distances (*md*) between the genotypes. The *md*-values ranged from 1.88 to 5.86 with mean 4.32.

Comparison of relationship measures

Simple (r) and rank (r_s) correlation coefficients between genetic similarity (*gs*) and 'well known' coefficient of coancestry (f_{wk}) were 0.404 (r) and 0.393 (r_s) within the core set of European two-row spring barleys (19 lines). The bootstrap 95%-confidence interval for r was [0.134, 0.642], indicating that r deviates significantly from 0. Including the six other barley lines from the core set, which had less-complete pedigrees, resulted in correlation coefficients 0.389 (r) , with bootstrap 95%-confidence interval [0.135,0.600] and 0.334 (r_s).

The correlation between *gs* and f_{wk} for the total set of barley genotypes (23 lines) including the North American and the winter barleys, is much higher. The value of r is 0.652 with bootstrap 95%-confidence interval [0.401,0.803] and the value of r_s is 0.711.

The relationship between *gs* and *f* can also be assessed by comparing the principal coordinate graph (Figure 2.1), based on the *gs* estimates, with the genotype pedigrees, in a more qualitative manner. The first observation that can be made is the clear separation of four of the six North American and winter barleys. Only the Canadian two-row spring variety Harrington and the German two-row winter variety Igri are positioned relatively close to the European two-row spring barleys. Harrington's pedigree, containing several European two-row spring barleys, confirms its position in the graph. Igri is more or less positioned in between the six-row winter variety Franka and a group of European two-row spring barleys. This is consistent with Igri's origin, i.e., a cross between a six-row winter barley hybrid and a two-row spring barley named Ingrid. The latter is also a parent of the cultivar Drossel which is positioned relatively close to Igri. Offspring of Emir, a Dutch two-row spring barley which was frequently used as a parent, appears to be concentrated in the front part of the graph: Aramir, Nudinka and Porthos. The Vada-, Aramir- and Isaria/Union/Volla/Trumpf-groups, as distinguished for the European two-row spring barleys by Melchinger et al. (1994), seem to emerge here as well. Furthermore, parent-offspring combinations are not very distant in the graph: Vada-Georgie, Kenia-Proctor, Aramir-Apex. However, the combination Apex-Riff seems to be rather distant. This picture may

be confirmed by the above average morphological distance of 4.59 between Apex and Riff and by personal communication with Dutch breeders who emphasised the clear agronomic differences between the two cultivars. It seems that selection against Apex-traits took place during the selection of Riff.

Simple (r) and rank (r_s) correlation coefficients between genetic similarity (*gs*) and morphological distance (*md*) were −0.124 and −0.142 within the core set of European two-row spring barleys for which morphological trait data were available (18 lines). The bootstrap 95%-confidence interval for r was [−0.362, 0.123], indicating no significant correlation between *gs* and *md*.

Correlation coefficients between f_{wk} and *md* could only be based on the thirteen lines that had 'well known' pedigrees, as well as morphological trait data, available. The value of r was −0.117 and the value of r_s was −0.189. From the bootstrap 95%-confidence interval for r [−0.363, 0.198] it was concluded that there was no significant correlation between f_{wk} and *md*.

Discussion

The degree of AFLP polymorphism does not appear to be very large in the set of barley genotypes we used. However, each primer combination set of markers could discriminate all genotypes. This may be a result of the choice of primer combinations which yield high numbers of unambiguous polymorphisms. It is not likely that a set of AFLP markers based on a randomly chosen primer combination will always be able to discriminate barley genotypes similarly well.

Although it does not have any direct effect on correlation estimates between genetic similarity (*gs*) and other relationship measures (f, *md*), it was decided to include monomorphic markers in the genetic-similarity estimation. One advantage of doing so is that the addition of extra genotypes in which a band of a so-far monomorphic marker is absent, making it a polymorphic marker, does not change 'existing' *gs*-estimates. If monomorphic markers are excluded, such an addition will result in a change of 'existing' *gs*-estimates. Similarly, by ignoring the simultaneous absence of a band in two genotypes, the addition of extra genotypes that have bands in 'new' positions will not change 'existing' *gs*-estimates.

The values of the correlation coefficients between genetic similarity and coefficient of coancestry are significant but not very high. This is in agreement with the poor-to-moderate correlations that were found between RFLP-based and RAPD-based *gs*-estimates and f (Graner et al., 1994; Tinker et al., 1993). One of the causes for this poor relationship may be inaccuracy in *gs*- and f-estimates.

The accuracy of *gs*-estimates depends on the number of markers, their distribution over the genome and the independent information (Messmer et al., 1993) provided by the AFLP-

markers. For the last reason, redundant markers with identical or allelic patterns within primer combinations have been discarded. Bootstrap analysis by sampling from all 681 markers resulted in standard deviation estimates for *gs* ranging from 0.006 to 0.012. An extra source of inaccuracy may be errors in scoring AFLP-bands. We tried to prevent part of these errors by scoring a data point as missing in case of doubt. The lack of information due to missing observations is included in the bootstrap standard-deviation estimates.

The assumptions underlying the calculation of coefficient of coancestry may cause quite some inaccuracy in f-estimation (Messmer et al., 1993). The assumption that original ancestors are equally unrelated with f=0 will probably not hold. It is quite likely that some pairs of ancestors, e.g. genotypes descending from the same region, are more related than others. Also the assumption that a genotype receives half of its genes from each parent is very doubtful. As a result of natural or breeder's selection during the inbreeding phase, alleles of one parent may have had the advantage over alleles of the other parent. As a result of this the estimated coefficients of coancestry may show substantial deviations from the true f-values.

The absence of a significant relationship between morphological distance (*md*) and *gs*- or f-estimates within the European two-row spring barleys may be a result of inadequate representation of genetic relationships by the observed morphological traits. Reasons for this could be: the limited number of traits observed, the limited variation for these traits, the number of underlying genes for these traits, which may also be limited, and possible epistatic interactions among these genes. Also the distribution of the underlying genes over the genome may be quite irregular. Finally, most data were measured on the rather coarse ordinal UPOV-scales (UPOV, 1981), which may have caused some inaccuracies in the *md* estimates. The poor within-group correlation can be said to agree with the results of Souza and Sorrells (1991b) in oats. The moderate correlation between gs and distance based on agronomic and morphological traits found by Autrique et al. (1996) in durum wheat is a result of the wider range of genotypes under investigation, representing more than one ecotype and resulting in much more variation among distance estimates. Also most of the observed traits were measured on a continuous scale, probably resulting in a higher accuracy of the distance estimates.

The accuracy of the correlation coefficient (r) estimates cannot be assessed straightforwardly, because the usual assumptions of independent samples of data-pairs from a bivariate normal distribution do not hold. The data-pairs, consisting of relationship measures, are dependent and have a non-normal distribution. In our case they are based on a genotype sample from the population of European two-row spring barleys. To avoid complex analytical approaches, bootstrap sampling from the genotypes can be used to approximate the proper confidence intervals for r. Inaccuracy appears to be larger than one would expect on the basis of the usual, but false, distributional assumptions. The addition of genotypes that did not have

'well known' pedigrees slightly decreased *r*, showing the effect of inaccuracy of *f* due to incomplete pedigree information.

Including genotypes from other barley groups, e.g. European winter barleys and North American spring barleys, resulted in a much larger estimate of *r*. The main reason for this bias is the simultaneous study of within- and between-group (*gs*,*f*) pairs. The higher value of *r* shows that AFLP-based *gs*-estimates can be used to distinguish between major groups of barley and suggests that genetic diversity in barley may very well be assessed with AFLPs.

The prediction of breeding behaviour of offspring from parent combinations may be improved by the simultaneous use of AFLP-based genetic similarities and coefficients of coancestry. A preliminary standardisation could be helpful in this respect to take account of the differing *gs* and *f* ranges. The combination of the *gs* and *f* estimates is expected to decrease the effect of their independent inaccuracies. The weights given to both relationship measures may depend on the number of markers and maybe the approximate inaccuracy of *f* (Cox et al., 1985a). However, the expected improvement of a combined measure can be made ineffective if *gs*- or *f*-estimates are biased (Souza and Sorrells, 1991b). Whether morphological distances have any predictive value on breeding behaviour remains questionable.

Conclusion

The AFLP fingerprint technique can be used for cultivar identification in barley. One primer combination may often be sufficient to identify lines uniquely.

Genetic similarities (*gs*), based on AFLP markers, show a poor-to-moderate correlation with pedigree based coefficients of coancestry (*f*) within the group of European two-row spring barleys. This poor relationship may be caused by inappropriate assumptions in the calculation of *f* as well as marker sampling error and biased representation of genomic differences revealed by AFLPs. Morphological distances (*md*) show no significant relationship with *gs* or *f*. This may be caused by biased and insufficient representation of the genome using morphological traits. The inaccuracy of the correlation coefficients between relationship measures, e.g. *gs*, *f* and *md*, can be assessed using bootstrap sampling of genotypes.

The clear distinction between major barley groups, based on *gs*-estimates, suggests opportunities for the use AFLP markers in the assessment of genetic diversity. For the prediction of breeding behaviour of parent combinations simultaneous use of several relationship measures (*gs*, *f*) in a combined index, as proposed by Cox et al. (1985a), may probably improve results if large biases in the *gs*- and *f*-estimates are absent. This improvement will be a result of the decreased effect of the individual inaccuracies.

Acknowledgements

The authors thank Izak Bos for critically reviewing the manuscript. They also thank all breeders and researchers who contributed pedigree information for this study.

3

Cross and line prediction in barley using F4 small-plot yield trials [2]

Johan W. Schut, C. Johan Dourleijn and Izak Bos

Abstract

Twenty crosses of European two-row barley were used to investigate the usefulness of F4 yield testing in small plots. We examined opportunities for selection between and within crosses. F2-derived F4 lines were tested at two locations in 1994 and the 48 descending recombinant inbred lines (RILs) per cross were tested for yield at two locations in 1995 and four locations in 1996. To rank crosses we used the cross prediction method proposed by Jinks and Pooni (1976, 1980), which uses a predicted mean and variance to estimate the probability that a line exceeds an arbitrary threshold yield. Correlation between F4 population mean and RIL mean was poor. This was mainly due the effect of intergenotypic competition between the small plots. Correlation between RIL mean and midparent value based on large plots was stronger, but the difference between RIL mean and midparent value varied significantly among crosses. This is probably a result of segregation distortion for 'major genes' and epistatic effects. The yield variance of F4 populations was moderately correlated with RIL variance. This can be mainly attributed to the presence or absence of segregating major genes in the different crosses. Cross prediction performed poorly, which is mainly due to the inadequate prediction of RIL mean by midparent values and RIL variance by F4 variance. Considering selection within crosses, we found that the yield of individual F4 lines did not accurately predict the yield of the related RILs. F4 yields were biased due to intergenotypic competition between the small plots. Also inaccuracies due to the small plot size contributed to the lack of prediction precision. In some crosses a significant relationship between F4 yield and RIL yield was established on the basis of variation caused by 'major gene' segregation. In most cases this variation can be assessed visually. So we conclude that there is hardly any perspective for a laborious early generation small plot yield assessment in practical barley breeding, neither for selection within crosses nor for selection between crosses.

[2]submitted for publication

Keywords: cross prediction, early-generation selection, *Hordeum vulgare*, major genes, segregation analysis

Introduction

In a breeding programme of a self-fertilising crop selection between crossing populations can be distinguished from selection within these populations. Part of the selection between crosses is already done before the actual crosses are made, by choosing only the most promising parent combinations. Final selection between crosses and selection within crosses are both performed during and/or after the inbreeding phase. In order to focus within-cross selection on the most promising crosses, Jinks and Pooni (1976) proposed a cross prediction method for selection between crosses, based on statistics derived from basic generations, i.e., F1, F2, B1, and B2 (Mather and Vines, 1952) or based on a triple test cross design, involving crosses of the F2 with both parents and the F1 (Kearsey and Jinks, 1968). Later they suggested prediction on the basis of F3-lines which reduces the amount of crossing labour and trait evaluation (Jinks and Pooni, 1980). This prediction is based on the mean performance of the F3-lines and the variance among them, assuming a normal distribution of the quantitative trait values. Crosses are ranked on the basis of the estimated probability of a line exceeding an arbitrarily chosen threshold, e.g. the performance of the best parent. After selection of the top-ranking crosses a line or pedigree selection method (Bos and Caligari, 1995) can be used to select within these crosses.

There are several reasons why selection in the early generations of inbreeding may be less effective than selection in advanced generations. Firstly, dominance, as well as epistasis involving dominance effects, has an adverse effect on the contrasts between genotypes. Secondly, in cereals like wheat and barley, the agronomic performance of a genotype can only be assessed by growing the material in plots. Intergenotypic competition within these plots appeared to be an important factor in decreasing the effectiveness of selection for yield in monoculture (Van Ooijen, 1989a,b). In addition, the small plot size in early generations, due to the large number of lines under investigation and the limited amount of available seed per line, results in interline competition between plots. This also substantially decreases the effectiveness of selection (Van Oeveren, 1992). Finally, genotype by environment interaction may cause undesirable crosses and/or lines being selected, as the number of early generation test locations and years is small, usually one or two locations and one or two years. This limited number of environments is also a result of large numbers of lines and few available seeds per line, and furthermore time pressure is limiting the number of test years.

These disturbing factors apply to the effectiveness of selecting promising lines within crosses, as well as to the power of assessing differences between crosses, in early generations.

However, several authors report reasonable results of cross prediction for yield in wheat (Snape, 1982) and barley (Tapsell and Thomas, 1983) by the use of the triple test cross design. The suggested use of F3- or doubled haploid lines (Jinks and Pooni, 1980) appears to be an efficient alternative for the triple test cross design (Snape, 1982; Caligari et al., 1985; Choo, 1988). Instead of estimating genetic parameters, Thomas (1987) shows that it is also possible to assess the expected number of desired genotypes per cross by simply counting the number of desired genotypes in the F3-generation. All of the mentioned authors, except Thomas (1987), derive statistics from basic generations tested in a chosen environment to predict cross performance observed in the same environment. However, the agreement between cross predictions based on early generation yield data from one environment and observed cross performance in another environment appears to be rather poor (Caligari et al., 1985; Thomas, 1987).

Within crosses the effectiveness of early generation selection for yield is questionable (Knott, 1979). Correlations between early and later generations are moderate (DePauw and Shebeski, 1973) or change from year to year due to genotype by environment interaction (Briggs and Shebeski, 1971). Sneep (1977) recommends unreplicated early generation trials with a high density of standard varieties (1:3). Only the central row of three-row plots should be harvested to avoid competition effects. However, Spitters (1979) argues that harvesting all three rows gives more reliable yield estimates as the enlarged precision due to the increased harvested area compensates for the inaccuracy due to interplot competition.

In this study we try to generalise the predictive value of early generation yield assessment by using a set of crosses larger than the sets used in earlier studies mentioned above. To investigate the effectiveness of cross prediction and within-cross selection, we used eighteen crosses of European two-row spring barley plus two reciprocals. Special attention is paid to the effect of segregating ‘major genes’. We will discuss the effects of interplot competition and genotype by environment interaction. Finally we will deliberate on the perspectives of early generation yield testing for practical breeding.

Material and methods

Plant materials

Eighteen spring barley (*Hordeum vulgare* L.) lines (Table 3.1a) were used as parents in a partial diallel crossing design (n=18, s=2; Kempthorne and Curnow, 1961) to produce 18 F2-populations (Table 3.1b). For two parental combinations reciprocal crosses were made, increasing the total number of crossing populations to 20. The parent lines were chosen to represent the germplasm employed in commercial two-row barley breeding programmes in

Table 3.1. Plant materials

a. Parent lines, their country of origin and year of release

genotype	country of origin	year of release	genotype	country of origin	year of release
Apex	the Netherlands	1983	Gunhild	Denmark	1980
Baronesse	Germany	1989	Karat	Czech Republic	1981
Bonaire	the Netherlands	1992	Midas	United Kingdom	1970
CEB-9079	the Netherlands	-	Porthos	France	1975
CEB-9186	the Netherlands	-	Prisma	the Netherlands	1985
Drossel	Germany	1971	Riff	the Netherlands	1993
Forester	United Kingdom	1990	Triangel	the Netherlands	1990
GEI-119	the Netherlands	-	Yriba	France	1981
Georgie	United Kingdom	1975	ZE-87-3414	the Netherlands	-

b. Crosses and their parents. R=reciprocal combination

cross	mother	father	cross	mother	father
1	Riff	Drossel	11	Karat	Yriba
2	Baronesse	Forester	12(R2)	Gunhild	GEI-119
3	Baronesse	Bonaire	13	Gunhild	CEB-9186
4	Apex	Riff	14	Bonaire	Porthos
5	Porthos	Yriba	15	CEB-9186	ZE-87-3414
6	Midas	Forester	16	ZE-87-3414	CEB-9079
7	GEI-119	Midas	17(R2)	GEI-119	Gunhild
8(R1)	Prisma	Apex	18	Triangel	Georgie
9	Prisma	Karat	19(R1)	Apex	Prisma
10	Triangel	Drossel	20	Georgie	CEB-9079

Northwest Europe over the last 20 years. Sixty F2-plants per cross were used to produce 60 F2-derived F4-lines. These 1200 F4-lines were used to obtain early generation yield data. From the same F2-plants 60 F5-plants per cross were derived via single seed descent. Out of these F5-plants we took a random sample of 48 plants per cross to produce, by one generation of multiplication, 48 F5-derived F7-lines per cross . These recombinant inbred lines, 960 RILs in total, were used to obtain estimates for final cross performance.

F4-trials 1994

The 1200 F4-lines were tested in 1994 at two clay-soil locations in the Netherlands (Lelystad: 94-1a, and Wageningen: 94-2a) in three-row plots of 1.5 m length. A few F4-lines were planted in only one or two rows due to lack of seed. The other rows were filled with the standard cultivar Magda. The distance between rows was 20.8 cm. The 18 parent lines plus two extra lines (Magda and Vada) were added as standards. Every F4-line occurred only once per location; standards were replicated six times, adding up to a total of 1320 plots per location. F4-lines were randomised according to a partially balanced incomplete block design with 10 plots per block, treating locations as replicates. One standard was added to every block, increasing the number

Table 3.2. Field trial description. p=present in trial

trial	location	year	plot size (m^2) (width × length (m))	sowing date	standards	F4	RIL
94-1a	Lelystad	1994	0.94 (0.625 × 1.5)	31 May	p	p	
94-2a	Wageningen	1994	0.94 (0.625 × 1.5)	1 June	p	p	
94-1b	Lelystad	1994	8.55 (1.5 × 5.7)	22 April	p		
94-2b	Wageningen	1994	8.55 (1.5 × 5.7)	28 April	p		
95-1	Swifterbant	1995	9.0 (1.5 × 6.0)	15 May	p		p
95-2	Wageningen	1995	9.0 (1.5 × 6.0)	11 May	p		p
96-1	Swifterbant	1996	9.0 (1.5 × 6.0)	1 April	p		p
96-2	Wageningen	1996	9.0 (1.5 × 6.0)	19 March	p		p
96-3	Lelystad	1996	5.32 (1.4 × 3.8)	18 April	p		p
96-4	Rilland	1996	3.6 (1.5 × 2.4)	19 March	p		p
96-5	Ottersum	1996	4.65 (1.5 × 3.1)	18 March	p		p

of plots per block to 11. Due to late ripening of F3-lines in the greenhouse, we sowed on 31 May in Lelystad and on 1 June in Wageningen. Although these dates were extremely late for Dutch conditions, crop development was generally normal as a result of cool weather during the first six weeks after planting. The three-row plots were harvested and the observed grain yield data were converted to kilogram per hectare.

Standard trials 1994

In 1994 the 18 parent lines plus the two extra lines (Magda and Vada) were also tested in large plots (10 rows; 1.5 x 5.7 m) at the same locations as the F4 (Lelystad: 94-1b, and Wageningen: 94-2b). We used two replicates per location and we applied a randomised complete block design. Sowing dates are presented in Table 3.2. Observed traits were plant height (cm), thousand kernel weight (g) and yield (kg/ha). Standard trials, as well as all other trials mentioned in Table 3.2, were kept free from diseases.

RIL-trials 1995 and 1996

The 960 recombinant inbred lines (RILs) were tested in 1995 at two locations (Swifterbant: 95-1, and Wageningen: 95-2). Again, the 18 parent lines and cultivars Magda and Vada were added as standards. Each location included two replicates. Each standard occurred six times per replicate, adding up to a total of 2160 (2 x 1080) plots per location. In 1996 the RILs were tested at five locations in the Netherlands: two 'complete' locations (Swifterbant: 96-1, and Wageningen: 96-2), containing all 960 lines, and three 'partial' locations (Lelystad: 96-3, Rilland: 96-4, and Ottersum: 96-5), each containing one third of the lines (i.e., 16 lines) of each cross. The 'complete' locations involved two replicates and six times 20 standards per replicate, just as the 1995-trials. The 'partial' locations involved two replicates with the 20 standards occurring

twice per replicate, adding up to a total of 720 (2 x 360) plots per location. All genotypes were randomised according to a partially balanced incomplete block design with eight plots per block. The constraint that two genotypes do not occur more than once together in the same block, extended over all year by location combinations. Due to lack of seed of some lines, we filled empty plots with the additional standard cultivar Reggae. Sowing dates and plot sizes are presented in Table 3.2.

Observed traits were grain yield (kg/ha), thousand kernel weight (g; 0% moisture), plant height (cm) and flowering time (°C.days). Flowering time was defined as the temperature sum from emergence to decimal stage 49 (Zadoks et al. ,1974). Grain dry matter content at harvest was measured on a 100g-sample. Lodging was scored on a scale from 0 (no lodging) to 5 (severe lodging) around decimal stage 69 (Zadoks et al., 1974), except for trial 95-2. In this trial lodging was observed at stage 83 as there was hardly any lodging in earlier stages. Potentially useful covariates were observed wherever it seemed relevant: weed cover (%); bird damage (0-2); percentage of clay in dry matter content samples; secondary tillering (0-2).

Statistical analysis

F4-trials

The F4-yield data were analysed using average information REML (Gilmour et al., 1995). The linear mixed model included fixed effects for location, standards, crosses, standard by location interaction and cross by location interaction. Also strips of adjacent incomplete blocks were included as fixed effects, nested within the location effects. The block effects were assumed random, as well as the line within cross effects. Whenever a hypothesis considered the specific 1200 lines that were present in the trial, the line effects were assumed to be fixed. Finally, error variances were allowed to be different at the two locations.

Standard trials and RIL-trials

Standard trial data were analysed per trait using a linear model including fixed effects for standards, locations and standard by location interaction. We used average height of the two adjacent plots as a covariate in the analysis of the yield data.

RIL-trial data were analysed per trait (yield, thousand kernel weight, plant height, flowering time and lodging) using average information REML (Gilmour et al., 1995). Analysis of the plot data was performed per year by location combination (environment), because an overall analysis appeared not feasible due to computational limitations. The linear mixed model included fixed effects for standards, crosses and strips of adjacent incomplete blocks. The block effects were assumed random, as well as the line within cross effects. In the analysis of residuals and whenever a hypothesis considered the specific lines that were present in the trial, the line

effects were assumed to be fixed. Several concomitant variables (sowing date difference because of rain during sowing, weed cover, bird damage, average plant height of the two adjacent plots, dry matter content of grains at harvest, percentage of clay in yield samples, secondary tillering, harvest date difference because of rain during harvest) were included, if significant (F-test; α=0.05), to decrease error variance in the analysis of thousand kernel weight (tkw) and yield data. For the analysis of lodging data we fitted a proportional odds model (McCullagh and Nelder, 1989).

To overcome computational limitations we performed a combined analysis over years and locations (environments) by using the least squares means for the lines as input data for an analysis of variance. The linear mixed model included fixed effects for standards, crosses, standard by environment interaction and cross by environment interaction. The line effect was assumed fixed or random depending on the hypothesis, as mentioned for the analysis of individual environments. Between line variances were estimated per cross.

Residual analysis was performed for each environment to trace outliers among the data. These observations were excluded from the final analyses. Observations for standard cultivars were considered to be outliers if the absolute value of their standardised residual was greater than 2.80 ($P<0.005$). For line observations we looked for a combination of a high absolute value of the standardised residual and an extreme trait value compared to the other lines from the same cross. These within-cross outliers among the lines were traced using least squares means for yield, thousand kernel weight, plant height, and flowering time from the variance analyses per environment. Tests for within-cross outliers were performed for each cross separately. Per trait a t-test and one of its robust versions (Rousseeuw and Van Zomeren, 1990) were used to find univariate outliers among the line means per trait. However, some within-cross outliers can only be distinguished by a multivariate test (Barnett and Lewis, 1994), taking into account observations for several traits from one environment or observations for one trait from several environments. We used Wilks' test for multivariate outliers (1963) and one of its robust versions, based on the minimum volume ellipsoid estimator (Rousseeuw and Van Zomeren, 1990). These tests are multivariate versions of the univariate tests for outliers. In the case of severely skewed data, even after the deletion of an outlying observation, the robust tests became unreliable and were not used. Finally, we defined a combined probability for Type 1-error: $P_c=1-(1-P_r)(1-P_{cr})$ where P_r follows from the standardised residual of the observation and P_{cr} from the smallest P of the univariate or multivariate tests for within-cross outliers. A P_c-value less than 0.06 was then used as an indicator to find and remove the outlying observations.

Segregation analysis

On the basis of parent information we expected several crosses to segregate for ‘major genes’, e.g. the *denso* gene (Haahr and Von Wettstein, 1976) and the *ert-g* gene (Thomas et al., 1984). Thomas (1987) suggested to investigate the effect of these ‘major genes’ on the effectiveness of cross prediction in more detail. To find out which lines obtained which ‘major gene’ allele, we used an approach based on mixture models (McLachlan and Basford, 1988). The model assumes a mixture of two multivariate normal distributions with mean vectors μ_1 and μ_2, equal variance-covariance matrices Σ and unknown proportions p_1 and p_2. We used robust M-estimators for estimation of μ_1 and μ_2 and Σ, as proposed by Maronna (1976) using Huber’s (1964) ψ-function in the weights. The model was fitted using the idea for an iterative EM-algorithm described by Jansen (1993). The algorithm was implemented in SAS-IML (SAS Institute Inc., 1989). The existence in a cross of a segregating ‘major gene’ was accepted on the basis of a likelihood ratio test (α=0.05), in which the unimodal model (H_0: no ‘major gene’ segregating) was tested against the bimodal model (H_1: ‘major gene’ segregating). Critical values for the likelihood ratio were obtained by Monte Carlo sampling from a multivariate normal distribution with a variance-covariance matrix based on the observed data. When a bimodal model was accepted, we refer to this as a 'major gene' segregating in the cross. Once a ‘major gene’ was accepted, this factor was used as a covariate in the mixture model, while looking for additional ‘major genes’ . This procedure was repeated until no further ‘major genes’ were detected.

The least squares means for the lines were used as input, while fitting the mixture model for each cross separately. Plant height, flowering time, lodging, thousand kernel weight and yield means from the combined analysis over environments were used one by one in a univariate mixture model. But we also used these means in combination in a multivariate mixture model. The third way of tracing segregating ‘major genes’, involved the combined use of least squares means from the different ‘complete’ environments (95-1, 95-2, 96-1, 96-2) for one trait at a time (plant height, flowering time, thousand kernel weight or yield) in a multivariate mixture model.

With the use of the postulated ‘major genes’, the genetic variance between lines could be divided into two parts: variance caused by the hypothesised ‘major genes’ and variance resulting from the segregation of other -‘minor’- genes. The latter variance is therefore called ‘minor gene’ variance. ‘Minor gene’ variance estimation is performed by using the segregating ‘major genes’ and their interactions as explanatory variables in the analysis of variance. In the analysis over environments we included the effect of ‘major gene’ by environment interaction. The postulated ‘major genes’ based on RIL data were also used in the analysis of the F4-lines, by assuming that the RIL and the F4-line descending from the same F2-parent, had the same 'major gene' genotype. Because 12 of the 60 F2-plants per cross were not advanced to recombinant inbred lines, we could not assign them a ‘major gene’ genotype. In the ‘minor gene’

Table 3.3. Field trials characterised by average trait values, observed over 20 crossing populations and 20 standards and root mean square errors (root mse) obtained by variance analysis. tkw = thousand kernel weight. †=average over 20 standards

trial	mean					root mse			
	plant length (cm)	flowering time (°C.days)	lodging (0-5)	tkw (g)	yield (kg/ha)	plant length (cm)	flowering time (°C.days)	tkw (g)	yield (kg/ha)
94-1a					2430				504
94-2a					3147				463
94-1b†	93			48.1	7081	4.4		1.6	347
94-2b†	81			49.8	5683	4.0		1.0	235
95-1	81	618	0.6		5452	3.1	9.7		229
95-2	77	593	1.5	46.4	6048	3.1	11.4	1.5	220
96-1	90	690	3.1	44.2		3.4	9.6	2.2	
96-2	80	643	2.7	48.0	9127	3.3	8.6	1.6	293
96-3	94		1.5		7404	2.8			301
96-4	87		2.8		9813	3.3			453
96-5	89		0.7		9067	2.9			261

variance analysis the F4-lines descending from these F2-plants were treated fixed instead of random. In this way they did not contribute to the estimated 'minor gene' variance.

Results

F4-trials 1994

In the small plot trials yield, averaged over the 20 F4-populations and 20 standards, was 2430 kg/ha in trial 94-1a and 3147 kg/ha in trial 94-2a (Table 3.3). Average yields per F4-population are presented in Table 3.4. The grain size of F4-seed, produced by the F3-lines in the greenhouse, was generally smaller than the grain size of the standards' seed, produced in the field. As we used a constant amount (weight) of seed for sowing, this caused large differences in plant density, resulting in lower yields for the standards than for the F4-lines in trials 94-1a and 94-2a. Estimated between-line standard deviations per F4-population for the combined locations are presented in Table 3.5. Lack of seed, and planting and harvest errors, resulted in 35 missing observations. Part of the plots in trial 94-1a were harvested later, due to rainy weather during harvest. Some of these plots suffered from severe pre-harvest sprouting. Yield of these plots was not observed, resulting in 13 extra missing observations. The effect of late harvest was included in the analysis by treating the plots in question as if they were from a separate location. The average yields for trial 94-1a in Table 3.3 were corrected for the effect of this late harvest.

Table 3.4. Average yield (kg/ha) per cross. (F4: 94-1a, 94-2a; RILs: 95-1,..,96-5)

	trial							
cross	94-1a	94-2a	95-1	95-2	96-2	96-3	96-4	96-5
1	2909	4038	5500	6218	9291	7443	9770	8968
2	3109	3814	5482	5999	9151	7788	9622	9102
3	2964	3836	5582	6354	9370	7518	10096	9295
4	3268	3952	5465	6211	9225	7479	9823	8850
5	2156	3076	5459	6019	8765	7174	9134	8758
6	2172	3463	4982	5870	8771	6652	9100	8727
7	2721	3982	5722	6493	8847	7039	9765	9438
8	2441	3106	5376	5766	8913	7544	9453	8899
9	1573	3011	5466	6112	9121	7712	9960	9686
10	2242	3922	5459	6265	9288	7270	9997	9442
11	1850	2880	5155	5825	8703	6909	9329	8988
12	3115	4205	5562	6342	9109	7270	9658	8845
13	2248	3466	5298	6087	8979	7273	9606	9144
14	2475	3516	5498	5897	9015	7237	9660	8721
15	2154	2953	5080	5652	8868	7331	9521	8472
16	2836	3094	5301	5579	8902	7377	9532	8776
17	3511	4298	5561	6358	9075	6933	9666	8882
18	2533	3588	5617	6111	9239	7324	10099	9140
19	2620	3281	5282	5739	8799	7547	9502	8866
20	2206	3440	5421	6285	9213	7274	9769	8867

Standard trials 1994

In the standard trials yield, averaged over the 20 standards, was 7081 kg/ha in trial 94-1b and 5683 kg/ha in trial 94-2b (Table 3.3).

RIL-trials 1995 and 1996

In 1995 yield, averaged over the 20 RIL populations and 20 standards, was 5452 kg/ha in trial 95-1 (Swifterbant) and 6048 kg/ha in trial 95-2 (Wageningen). We found 30 yield outliers by residual analysis, which were declared missing.

In 1996 average yields were higher, because of early sowing and cool weather during grain filling. Yields were 9127 kg/ha in Wageningen (96-2), 7404 kg/ha in Lelystad (96-3), 9813 kg/ha in Rilland (96-4) and 9067 kg/ha in Ottersum (96-5). Due to severe hail storm damage, we did not obtain yield data from Swifterbant (96-1). We could, however, sample spikes from this location to observe thousand kernel weights. Part of the Wageningen plots were harvested later due to rain. Severe pre-harvest sprouting in these plots resulted in 135 missing observations. Strong contamination with clay caused another 77 missing observations at the Wageningen location. By th use of residual analysis we found 11 yield outliers, which were declared missing.

Using within-cross residual analysis we also found five lines which consistently differed from the rest of the crossing populations to which they belonged. Assuming that these lines were

Table 3.5. Square root of between line variance for yield (kg/ha) per cross per environment (RILs: 95-1,..,96-5) and over environments (F4: 94-1a,2a; RILs: all-95,96)

	trial							
cross	94-1a,2a	95-1	95-2	96-2	96-3	96-4	96-5	all-95,96
1	328	342	460	587	311	514	904	397
2	258	161	330	528	426	522	629	292
3	227	225	317	463	327	420	371	252
4	255	249	315	504	168	431	516	254
5	336	206	248	257	252	527	405	197
6	431	461	448	620	310	708	651	461
7	193	300	360	343	314	411	346	232
8	345	196	337	595	339	355	653	357
9	248	215	292	313	278	146	296	147
10	261	293	232	258	353	370	263	199
11	426	314	417	431	491	649	619	393
12	366	285	273	222	520	365	520	153
13	315	252	229	357	282	319	505	234
14	303	227	452	419	489	528	752	356
15	650	244	396	526	390	405	646	289
16	405	314	406	488	294	501	380	336
17	246	353	311	253	470	416	452	217
18	333	238	299	274	537	175	494	241
19	347	213	303	568	464	530	714	344
20	370	351	380	570	459	628	555	492

products of cross pollination in early generations or accidental exchanges of genotypes, we treated them as additional standards in further analysis.

Average trait values and root mean square errors per environment are presented in Table 3.3. Average yields per cross are presented in Table 3.4. Square roots of the estimated between line variances per cross are presented in Table 3.5.

Segregation analysis

As a result of segregation analysis we were able to postulate 0 to 4 segregating 'major genes' per cross (Table 3.6). Whenever we expected segregation of the *denso* or the *ert-g* gene in a cross, on the basis of the pedigrees of the parents, we were able to identify this segregation by visual inspection of the distribution as well as by using the mixture model. Segregation ratios appeared to be significantly distorted in 25 of the 43 postulated 'major genes' (χ^2-test; α=0.05). Among the genes with distorted segregation was the *ert-g* gene, of which the erect allele occurred with a frequency of 0.30.

The postulated 'major genes' and their interactions explained 0 to 100 % of the 'total' yield variance, as well as 0 to 100% of the thousand kernel weight (tkw) variance, in the

Table 3.6. Number of postulated segregating 'major genes' per cross and the relative change of variance resulting from including them as explanatory variables in the analysis of variance over environments. Known 'major genes' that are segregating, are specified: *denso* (Haahr and Von Wettstein, 1976) and *ert-g* (Thomas et al., 1984). tkw=thousand kernel weight

cross	number of 'major genes'	known 'major genes' segregating	relative change in variance		
			F4-yield	RIL-yield	RIL-tkw
1	2	*denso*	-0.01	-0.34	-0.41
2	4	*denso*	-0.59	-0.57	-0.86
3	1		0.00	-0.04	-0.42
4	2	*denso*	-0.24	-0.54	-1.00
5	0		0	0	0
6	3	*denso, ert-g*	-0.18	-0.66	-0.71
7	3		-0.23	-0.33	-0.44
8	2	*denso*	-0.11	-0.56	0.04
9	0		0	0	0
10	2		0.05	-0.16	-0.64
11	2	*denso*	-0.01	-0.8	0.08
12	3	*ert-g*	-0.56	-0.25	-0.73
13	2		-0.51	-0.34	-0.22
14	4		-0.21	-0.30	-0.67
15	3		-0.06	-0.19	-0.53
16	1	*denso*	-0.04	-0.37	-0.02
17	3	*ert-g*	-0.40	-0.36	-0.67
18	2		0.00	-0.05	-0.04
19	2	*denso*	-0.68	-0.73	-0.06
20	2	*denso*	-0.28	-0.73	0.06

individual environments. Estimated over environments, 'minor gene' variances for yield or tkw were usually smaller than 'total' variances (Table 3.6). In a few cases slightly larger 'minor gene' variances were found. They result from crosses in which 'major genes' predominantly explain line by environment interaction variance. As a result of the separation of this interaction, the average line effect over environments can be estimated more accurately, leading to an increased between line variance over environments.

Cross prediction. 1. Mean

Yield of recombinant inbred line (RIL) populations, averaged over environments, was only moderately correlated (r=0.42) with yield of F4 populations (Table 3.7a). Midparent yield, calculated as the mean yield of the two parents of a cross, and based on small plot yield data from the same two trials as the F4, showed a similar correlation (r=0.45) with yield of the RIL populations. However, midparent yield, based on earlier sown, large plots at the same locations, showed a much higher correlation (r=0.70). This correlation is about equal to the correlation between RIL population yield and midparent yield, based on large plots in the same environ-

Table 3.7. Correlation coefficients between yield of crosses, averaged over recombinant inbred lines (RILs), and predictors based on F4 or parent data.

a. Yield (y) performance. y(cr)••=cross yield performance averaged over years *i* (*i*=1995, 1996) and locations *p* (p=1,..,5); y(cr)*ip*=cross yield performance in year *i* at location *p*; y(cr)*94a*•=cross yield performance (F4) averaged over trials 94-1a and 94-2a; y(mp)*94a*•= midparent yield averaged over trials 94-1a and 94-2a; y(mp)*94b*•=midparent yield averaged over trials 94-1b and 94-2b; y(mp)••=midparent yield averaged over years *i* and locations *p*; y(mp)*ip*, y(mp)*iq*, y(mp)*jq*=midparent yields in year *i* or *j* (*i*≠*j*) at location *p* or *q* (*p*≠*q*). †=correlation coefficients averaged over all possible combinations of *i*, *j*, *p* and *q* (*i*≠*j*; *p*≠*q*)

average cross	predictors of cross yield (y)						
yield	y(cr)*94a*• (F4)	y(mp)*94a*•	y(mp)*94b*•	y(mp)••	y(mp)*ip*	y(mp)*iq*	y(mp)*jq*
y(cr)•• (RILs)	0.42	0.45	0.70	0.71		0.45†	
y(cr)*ip* (RILs)	0.35†	0.36†	0.53†	0.54†	0.65†	0.37†	0.20†

b. Thousand kernel weight (k) per cross, averaged over RILs. n.a.=not available: thousand kernel weight not observed in trials 94-1a and 94-2a. Further descriptions as in a.

average cross	predictors of cross thousand kernel weight performance (k)						
k	k(cr)*94a*• (F4)	k(mp)*94a*•	k(mp)*94b*•	k(mp)••	k(mp)*ip*	k(mp)*iq*	k(mp)*jq*
k(cr)•• (RILs)	n.a.	n.a.	0.82	0.90		0.87†	
k(cr)*ip* (RILs)	n.a.	n.a.	0.81†	0.89†	0.88†	0.85†	0.86†

ments as where the RILs were tested (*r*=0.71; Figure 3.1). Midparent values, based on a single environment, showed only moderate correlations with RIL population yield (average *r*=0.45). For thousand kernel weight correlations between midparent values and RIL averages were higher (Table 3.7b).

Figure 3.1 shows that the difference between midparent yield and cross yield seems to vary among crosses. In most crosses the average yield of the RILs is not as high as predicted by midparent values, while for some other crosses midparent values and average yield of RILs seem to be equal. A nonparametric Friedman test, using environments as blocks, showed significant variation among crosses for the difference between midparent value and average yield of RILs (P=0.003). For thousand kernel weight the RIL average of each cross seems to be equal to the midparent value of that cross.

Cross prediction. 2. Variance

Yield variance between F4-lines in a cross showed a reasonable correlation (*r*=0.62; Table 3.8a; Figure 3.2a) with yield variance between recombinant inbred lines, estimated over environments. 'Minor gene' yield variance between F4-lines showed a weaker relationship (*r*=0.41; Table 3.8b; Figure 3.2b) with 'minor gene' yield variance between RILs. The extremely large yield variance

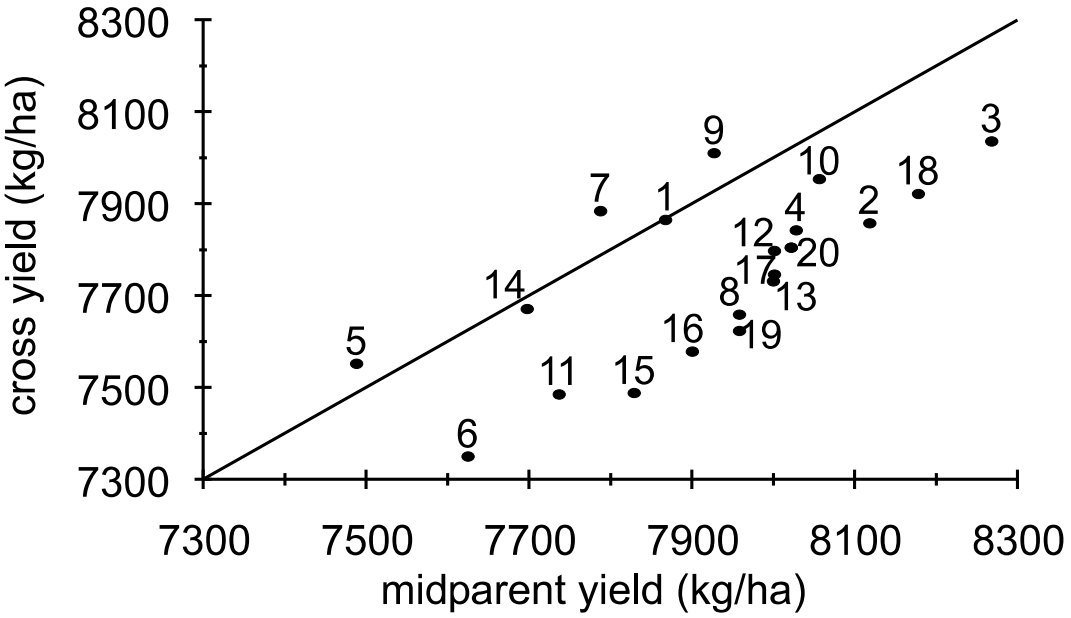

Figure 3.1. Yield of crosses (RIL-populations) and mean yield of their parents, both averaged over 6 environments in 1995 and 1996. The line represents the 1:1 ratio.

between F4-lines of cross 15 (see Table 3.5) was not used in the correlation coefficient estimation as some of the F4-lines hardly produced any seed due to late flowering. This was the result of late sowing in combination with the segregation of a photoperiod response gene and a late flowering parent (CEB-9186). The average correlation between RIL yield variance from a single environment and RIL yield variance estimated over the other environments is moderate (r=0.50; Table 3.8a) and decreases to 0.25 (Table 3.8b) when the effect of 'major genes' is removed. Correlations between RIL variances for thousand kernel weight are much higher (Table 3.8a), but they also decrease after elimination of 'major gene' effects (Table 3.8b).

The poor correlation between yield variance between F4-lines and thousand kernel weight variance between RILs (r=0.32; Table 3.8a) increases to 0.46 (Table 3.8b) when only 'minor gene' variance is considered.

Cross prediction. 3. Combining mean and variance

Cross performance was predicted by the estimated probability that a line from a certain cross yields more than an arbitrary threshold. The threshold was defined as the average of three high yielding standard cultivars: Riff, Baronesse and Triangel. The probability was calculated on the basis of a predicted mean, a predicted variance and on the assumption that line yields of a cross are normally distributed. For the RILs we observed that the standard deviation for yield, averaged

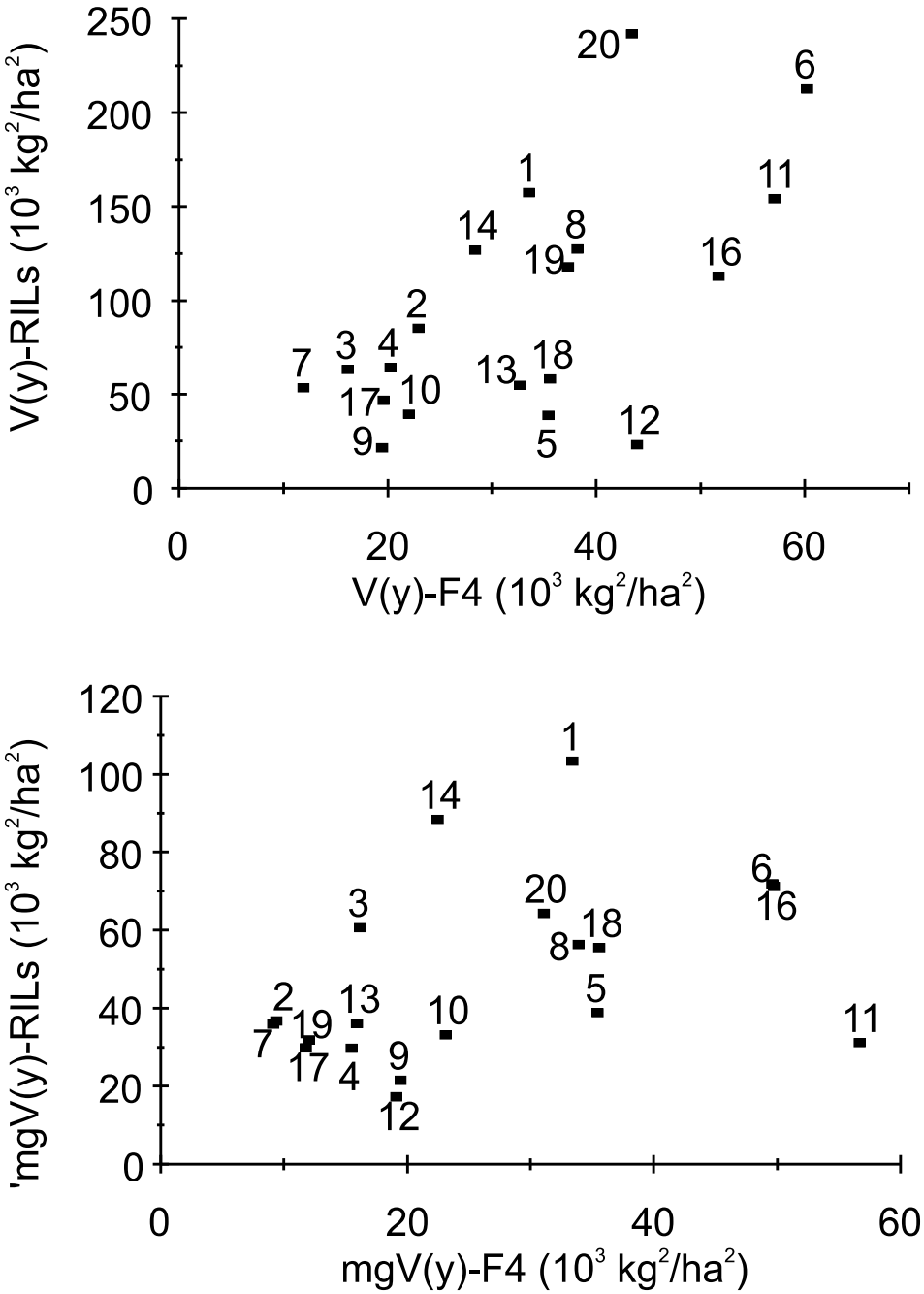

Figure 3.2. F4-line yield variance per cross, analysed over trials 94-1a and 94-2a, and recombinant inbred line (RIL) yield variance per cross, analysed over 6 environments in 1995 and 1996. Cross 15 not included.

a. Between line variances for F4-lines (V(y)-F4) and RILs (V(y)-RILs)

b. 'Minor gene' between line variances for F4-lines (mgV(y)-F4) and RILs (mgV(y)-RILs)

Table 3.8. Correlation coefficients among 'between line' variances of crosses, based on F4-lines or recombinant inbred lines (RILs), for yield (y) and thousand kernel weight (tkw).

a. Variances (V) including the effect of segregating 'major genes'. V(y)••='between RIL' variance for yield analysed over years *i* (*i*=1995, 1996) and locations *p* (*p*=1,..,5); V(y)*ip*, V(y)*jq*='between RIL' variance for yield in year *i* or *j* at location *p* or *q*; V(tkw)••='between RIL' variance for thousand kernel weight analysed over years *i* (*i*=1995, 1996) and locations *p* (*p*=1, 2); V(tkw)*ip*, V(tkw)*jq*='between RIL' variance for thousand kernel weight in year *i* or *j* at location *p* or *q*; V(y)*94a*•='between F4-line' variance for yield analysed over trials 94-1a and 94-2a; ‡='between F4-line' variance of cross 15 not included; †=correlation coefficients averaged over all possible combinations of *i*, *j*, *p* and *q* (*ip*≠*jq*); #=same as for †, except for restrictions on *i*,*j*,*p* and *q*; ##=same as for #, but V(y/tkw)•• (in r(var(y/tkw)*ip*,var(y/tkw)••)) analysed over all year by location combinations without *ip*

	V(y)*94a*• (F4)‡	V(y)•• (RILs)	V(y)*jq* (RILs)	V(tkw)•• (RILs)	V(tkw)*jq* (RILs)
V(y)•• (RILs)	0.62				
V(y)*ip* (RILs)	0.42	0.50##	0.34†		
V(tkw)•• (RILs)	0.32	0.40	0.43		
V(tkw)*ip* (RILs)	0.33	0.35	0.35#	0.84##	0.78†

b. 'Minor gene' variances (mgV). Further descriptions as in a.

	mgV(y)94a• (F4)‡	mgV(y)•• (RILs)	mgV(y)*jq* (RILs)	mgV(tkw)•• (RILs)	mgV(tkw)*jq* (RILs)
mgV(y)•• (RILs)	0.41				
mgV(y)*ip* (RILs)	0.24	0.25##	0.12†		
mgV(tkw)•• (RILs)	0.46	0.49	0.33		
mgV(tkw)*ip* (RILs)	0.36	0.39	0.25#	0.66##	0.56†

over crosses, varied over environments (Table 3.5). We established a linear relationship between yield (y_{ij}) in environment *j*, averaged over standards *i'* (18 parents, Magda, and Vada), and the RIL standard deviation for yield ($\sqrt{V(y)_{ij}}$), averaged over crosses *i*, in the same environment:

$$\sum_{i=1}^{20} \sqrt{V(y)_{ij}} \;/\; 20 \;=\; 0.050 \;\Big(\; \sum_{i'=1}^{20} y(standard)_{i'j} \;/\; 20 \;\Big) \qquad (R^2 = 0.99)$$

We could then calculate the predicted RIL variance for cross *i* in environment *j* as:

$$V(y)_{ij}^{predicted} \;=\; \Big[\; \frac{\sqrt{V(yF4)_i}}{\sqrt{V(yF4)}}\; 0.050 \;\Big(\; \sum_{i'=1}^{20} y(standard)_{i'j} \;/\; 20 \;\Big)\; \Big]^2$$

where $V(yF4)_i$ is the variance for yield between F2-derived F4 lines for cross *i* (Table 3.5), which is approximately half the additive genetic variance (Jinks and Pooni, 1980). The calculated cross predictions were compared with the observed cross performance, i.e., the observed frequency of recombinant inbred lines (RILs) outyielding the given threshold (Table 3.9). Correlations

between observed and predicted cross performance were virtually absent when using a predicted mean based on the small plot trials in 1994, either midparent value or F4-population yield. Also directly observed frequencies of F4-lines outperforming the average of the three standards in the small plot trial, as proposed by Thomas (1987), did not show any relationship with the observed frequencies in later generations. The use of a predicted mean based on midparent values of large plot trials in six environments increased the average rank correlations between observed and predicted cross performance to 0.22.

Table 3.9. Spearman rank correlation coefficients between predicted and observed cross performance. Cross performance is defined as the probability (P) of a line yield (y) exceeding a certain threshold yield (y(th)). The threshold yield is calculated as the average of the standards Riff, Baronesse and Triangel. Whenever predictions are based on data from 1995 or 1996 trials, they are compared with observations from the same environment or set of environments. Cross 15 is not used in the correlation coefficient estimation. P(obs)=observed probability; y(RIL)••=line yield averaged over years *i* (*i*=1995, 1996) and locations *p* (*p*=1,..,5); y(RIL)*ip*=line yield in year *i* at location *p*; y(F4)*94a*•=F4-line yield averaged over trials 94-1a and 94-2a; y(th)*ip*, y(th)••, y(th)*94a*•=threshold yield, based on the same trials as the line yields y(RIL)*ip*, y(RIL)••, and y(F4)*94a*•, respectively; †=correlation coefficients averaged over all years *i* and locations *p*. Further descriptions for predictors of mean as in Table 3.7, and for predictors of variance as in Table 3.8.

predictor of mean	y(mp)*94a*•	y(cr)*94a*•	y(mp)••	y(cr)••	y(mp)••	y(cr)••	P(obs) y(F4)*94a*• > y(th)*94a*•
predictor of variance	V(y)*94a*•	V(y)*94a*•	V(y)*94a*•	V(y)*94a*•	V(y)••	V(y)••	
P(obs) y(RIL)•• > y(th)••	0.00	0.14	0.22	0.48	0.46	0.80	-0.07
P(obs) y(RIL)*ip* > y(th)*ip*	0.10†	0.12†	0.22†	0.34†	0.22†	0.35†	0.09†

Line prediction

Observed over all crosses, the correlation between yields of an F4-line and a recombinant inbred line (RIL), derived from the same F2-plant, appeared to be rather weak and generally negative (Table 3.10). It also appeared that the magnitude and direction of this correlation varied significantly among crosses. Rank correlations between RIL plant height and F4 yield were generally positive (r_s=0.16), while correlations between RIL plant height and RIL yield were generally negative (r_s=−0.39). Overall rank correlations between RIL flowering time and F4-

Table 3.10. Spearman rank correlation coefficients (r_s) between yield of an individual F4-line (y(F4)••), and yield of the recombinant inbred line (y(RIL)••), derived from the same F2-parent as the F4-line. Correlations between yield residuals (y'), i.e., the part of yield variation that cannot be explained by segregation of 'major genes' , are also presented. F4-line yields are averaged over trials 94-1a and 94-2a. RIL yields are averaged over all trials in 1995 and 1996. *=0.10<P<0.05; ***=P<0.01

cross	r_s(y(F4)••,y(RIL)••)	r_s(y'(F4)••,y'(RIL)••) no 'major gene' effects
1	0.02	0.24
2	0.08	0.09
3	0.04	0.05
4	-0.28*	-0.03
5	0.11	0.11
6	-0.02	-0.12
7	0.05	0.05
8	0.12	0.23
9	0.00	0.00
10	0.05	0.09
11	-0.04	-0.14
12	-0.09	0.17
13	-0.27*	0.04
14	-0.25*	-0.25*
15	-0.08	-0.08
16	-0.26*	-0.25*
17	-0.21	0.06
18	0.00	0.09
19	-0.25*	0.05
20	-0.44***	-0.08
all crosses	-0.09***	0.01

yield were somewhat negative (r_s=−0.09), while correlations between RIL flowering time and RIL yield were generally positive (r_s=0.25). After removal of 'major gene' effects yield of F4-lines showed no significant correlation with yield of related RILs, when considered over all crosses. By looking at individual crosses we observed a general tendency towards more positive correlations than when including 'major gene' effects, but correlations were generally poor. The correlation of yield with RIL plant height moved towards zero after removal of the 'major gene' effects, with r_s=0.06 for F4-lines and r_s=−0.21 for RILs. This was even more the case for the correlation of yield with RIL flowering time, with r_s=−0.03 for F4-lines and r_s=0.02 for RILs.

Discussion

Although the early generation small plot yield trials had rather high CVs (0.21 for 94-1a and 0.15 for 94-2a), average yield of F4-populations corresponded well with midparent values derived from the same trial (r=0.87). The correlation of these F4-population yields with midparent values, based on the large plot yields grown in the same year, was moderate (r=0.52). Correlation with yield of RIL populations, averaged over 1995 and 1996 trials, was only a little weaker (r=0.45). Predicting RIL-population yields on the basis of F4-populations or parents grown in small plots (94-1a and 94-2a) appears not very reliable. Midparent values based on large plot trials in the same year (94-1b and 94-2b) give a much better prediction of the yields of the RIL populations (r=0.70). The predictive value of these midparent values approaches that of midparent yields obtained from the same trials as the RIL-population yields (r=0.71). Using midparent values from individual environments to predict population yields over all environments is clearly more difficult (r=0.45), mainly because of genotype by environment interaction.

However, midparent values seem to have a limitation in their predictive value for RIL-population yields, even when these yields are averaged over six environments. There appears to be a substantial difference in yield level between the parental average and their offspring in a large proportion of the crosses (Figure 3.1). This may be explained by epistatic effects, where favourable combinations of genes in the parents are lost in most of the offspring. Evidence for this type of locus interaction was found in rice by Li *et al.* (1997). Another explanation could be distorted segregation of 'major genes' for yield. In most crosses with a large difference between midparent and offspring yield, we found that the skewed segregation ratios for the postulated 'major genes' could indeed explain the lower yield of the RILs. It may be questioned, however, whether the postulated segregating 'major genes', which are based on phenotypic observations, express the effect of segregation at a single locus. They may also be the result of two segregating loci with epistatic effects, which supports again the first explanation. Some evidence for both hypotheses was found, using molecular marker information from the two pairs of reciprocal crosses. The segregation distortion for the *ert-g* gene was confirmed in crosses 12 and 17 (Koorevaar, unpublished). However, segregation distortion for a postulated 'major gene' in crosses 8 and 19 (not *denso*), could not be confirmed by a single segregating marker (Yin, unpublished). Therefore, we conclude that segregation distortion as well as epistasis can cause decreased average yields of offspring compared to their parents.

The F4-line based prediction of yield variance between RILs is less reliable than predicting the mean. This is mainly due to the estimation errors in the variance components, caused by the relatively large error variance in the F4 and the limited number of tested lines. Estimation error for the RIL variances also had a disturbing effect on the correlation coefficient

estimates. This effect, however, is not relevant in the practical prediction situation. We can conclude that estimated variances based on a single environment, lack the accuracy to be good predictors of yield variance over environments, as the average correlation coefficient between them was only 0.50. Small plot sizes do not seem to affect the variance predictions very much as the average correlation coefficient of the single-environment based variances with the F4-variance is 0.42, which is not much less than 0.50.

However, the moderate correlations between F4-line variance and RIL variance seem to be mainly a result of segregating ‘major genes’ . This causes large line differences in yield within part of the crosses, while other crosses lack this variation. Removing the effect of the 'major genes' results in a large decrease of the correlation coefficient between yield variances (Table 3.8b). This corroborates the statement of Thomas (1987), that cross prediction should be performed separately for crosses with a segregating *ert-g* or *denso* gene and crosses without this segregation. Most of the studies on cross prediction, cited in the introduction, do not mention the effect of segregation of ‘major genes’, although this may very well have contributed to the positive results of cross prediction that were described. Excluding ‘major gene’ effects from the variance estimations also shows an increased correlation between thousand kernel weight variance and yield variance (Table 3.8a,b). If we assume that 'minor gene' variance is a joint result of the segration of many genes with small effects, the magnitude of this variance is linearly related to the number of segregating genes. This relationship will be trait-independent. An increased correlation between 'minor gene' variances for yield and thousand kernel weight supports the assumption that 'minor gene' variance of a trait is a result of the fraction of the total genome that is segregating in the cross. As parental relatedness is assumed to be linearly related to the number of segregating genes (see chapter 2), it may be used to predict 'minor gene' variance.

It is questionable whether in practical breeding the high correlations between F4-line variance and RIL variance, due to ‘major gene’ segregations, are relevant. First, practical breeders will often know when to expect a segregating ‘major gene’ like *ert-g* or *denso*. They will select visually for the desired allele in early generations, so that the variation caused by segregation has already disappeared in the final yield trials. In these later generations only 'minor gene' variance can be exploited, for which we showed that its prediction, on the basis of F4-line data, is difficult. Possibly a prediction on the basis of relationship measures of the two parents (see chapter 2) is more accurate. Secondly, in practical breeding ‘major gene’ segregation will not occur as frequently as in our material, as breeders try to avoid this segregation by making crosses between more related genotypes. However, one can envisage situations in which the prediction of ‘major gene’ segregation is very relevant for practical breeding. This is the case when potentially promising parent combinations are chosen on the basis of QTL-analyses, in combination with pedigree information, so that one may expect segregation at certain QTL.

Knowledge of the expected segregation patterns is useful to decide upon the necessity of 'major gene' selection in early generations and the number of offspring lines grown in these generations. And, of course, combining alleles of 'major genes' from both parents may result in transgressive offspring lines, that show clear improvement in trait value in comparison with both parents.

The prediction of cross performance gave poor results (Table 3.9). Mean prediction on the basis of midparent values from large plots over several environments, in combination with F4-based variance prediction produced the 'best' predictions (r_s=0.22). An 'exact' prediction of the cross mean, based on RIL yields, would have improved cross prediction quite significantly (r_s=0.48). An 'exact' variance prediction, based on between RIL variance, would have resulted in an almost similar improvement of the prediction result (r_s=0.46). However, predicting cross performance in individual environments, although difficult, can seemingly only be improved by accurate mean predictions. The rank correlation of 0.80, between 'exact' prediction of cross performance and observed cross performance shows that the observed number of lines outyielding the threshold is also subject to random error. Improvement of this ' exact' prediction by adding information about 'major gene' segregation ratios and allele effects, in combination with 'minor gene' variance, hardly gave any improvement in the correlation with observed cross performance (r_s=0.81).

Considering the poor correlations between yield of an individual F4 line and yield of a related RIL, we conclude that perspectives for small plot yield assessment of individual F4-lines are very limited. Differences between crosses are relatively large and yield correlations are often due to segregating 'major genes'. As mentioned above, the desired alleles of these 'major genes' can often be assessed visually instead of by a labourious yield assessment. This may very well explain the positive results of early generation within-cross yield selection in small cereals, reported by several authors (e.g., DePauw and Shebeski, 1973). The poor correlations that we found are partially due to the inaccuracy of F4-line yield assessment with only one replicate per environment and rather high CVs. The latter may be a result of the small size of individual plots with insufficient compensation for small-scale soil irregularities. Also line by year and line by location interaction cause a decrease in correlation between F4-line and RIL yield: average rank correlation between RIL yields within crosses from different environments was only 0.35 and decreased to 0.30 after elimination of 'major gene' effects.

Correlation between yield of an F4-line and the yield of a recombinant inbred line descending from the same F2-plant was generally negative. The negative correlation is probably related to the effect of plant height. In small F4-plots tall genotypes are at an advantage as a result of strong interplot competition. In the large RIL plots plant height is negatively correlated with yield, because tall plants lodge more often and have a reduced harvest index. In this study the effect of small plot size is somewhat confounded with the effect of late sowing so that it is

not completely clear which part of the inaccuracy is caused by the plot size. As already mentioned above, by considering RILs, instead of F4 lines, we observe a large change in overall rank correlation between yield and plant height, from positive to negative. The analogous change in overall rank correlation between yield and flowering time, from negative in the F4, to positive in the RIL populations, is less pronounced. Some F4-lines were too late under this late sowing to produce high yields. Also the segregating *denso* gene in several crosses is causing this negative correlation between flowering time and yield in the F4. The mutant allele causing shorter plants also has an effect towards lateness. So plants possessing this allele do not yield less because of late flowering, but because of their short stature. This explains the disappearance of the correlation between flowering time and yield after the elimination of 'major gene' effects. Altogether the main disturbing factor in the F4 trials appears to be the intergenotypic competition between the small plots, confirming the conclusion of Van Oeveren (1992). Whether intergenotypic competition within plots has any effect on early generation yield trials cannot be said on the basis of this study.

Conclusion

Considering the results of this investigation we conclude that prediction of cross yield performance on the basis of small plot yield trials in the F4 is not very reliable. We used two locations to decrease the effect of cross by location interaction. Intergenotypic competition between the small plots results in biased estimates of line, cross, and standard cultivar yields. But even prediction of cross population mean by midparent values from large plots, averaged over a wide range of environments, cannot improve reliability to a level that shows perspectives for practical breeding. Segregation distortion and epistatic effects of yield genes may cause differences between midparent values and cross means, differences varying from cross to cross.

The prediction of genetic variance for yield between recombinant inbred lines is based on yield variance between F4-lines. Although variance estimates are not very accurate, the correspondence between the two variances seems reasonable. However, it appears to predict mainly differences between crosses with respect to the numbers of segregating 'major genes' and the magnitude of their effect on yield. Some of these 'major genes' do not regularly segregate in commercial breeding programs and if they segregate, they are usually selected visually in early generations. This will mainly result in RIL populations without segregating 'major genes' , where only 'minor gene' variance can be exploited. Unfortunately 'minor gene' variances in F4-populations do not reliably predict 'minor gene' variance in RIL populations.

Prediction of individual recombinant inbred line yields within crosses on the basis of the yield of the F4-line descending from the same F2-plant, is not feasible. This is due to the large

inaccuracy of small plot trials, intergenotypic competition between plots, line by location interaction and line by year interaction. Also the genotypes of the RIL and the F4 line are not completely the same, although they were derived from the same F2-plant.The main correspondence between an F4-line and its inbred descendant is based on the 'major gene' constitution of the F2. As it is often possible to assess this visually, it is not necessary to perform an accurate yield assessment to select between lines within a cross.

Provided that the amount of seed is not limiting, the use of larger plots may improve line prediction within crosses, as well as cross mean prediction. However, it is highly questionable whether such an improvement is large enough to compensate for the cost of early generation yield assessment.

Acknowledgements

The authors thank Cebeco Zaden B.V., Vanderhave Research and Zelder B.V. for making available some of their advanced breeding lines and extensive trial facilities. They thank CPRO-DLO and CIMMYT, Mexico for the use of trial facilities. They also thank Piet Stam, Ritsert Jansen, Fred van Eeuwijk, Johannes Kalma, Derk Gesink, XinYou Yin and Gerard Koorevaar for their contributions to this study.

4

Prediction of progeny variation in barley crosses using parental relationship measures. I. Measures based on pedigree, morphological, agronomic or AFLP data [3]

Johan W. Schut and Izak Bos

Abstract

Twenty two-row spring barley crosses, each represented by 48 recombinant inbred lines, were used to study the prediction of progeny variance using several parental relationship measures. These distances were based on pedigree ($1-f$), morphological (*md*), agronomic (*agd*) and AFLP ($1-gs$) data separately, or in combination. No significant correlations were found between *md* and progeny variance for the four investigated traits, i.e., plant height, flowering time, thousand kernel weight and grain yield. Also for $1-f$ no significant correlations were found, although they were generally positive. However, only 10 crosses could be used due to lack of reliable pedigree data. Agronomic distances (*agd*), based on parental traits showed several positive and significant correlations with progeny variance, especially when parent traits for *agd* estimation were the same as the RIL traits for variance estimation. However, the associations appeared to be mainly based on differences in 'major gene' effects between crosses. In most breeding populations progeny variation based on 'major genes' is absent due to early generation selection or absence of allelic differences between parents. Thus *agd* is not expected to be a succesful predictor of progeny variance in practical plant breeding. The correlations between $1-gs$ and progeny variance were mainly positive, but insignificant. Correlation coefficients did not show consistent differences when considering total and 'minorgene' variance. The lack of association is expected to be a result of lack of representation of QTL controlling the investigated traits by the observed AFLP markers. In general combined distance estimates showed the highest correlations with progeny variance. A distinction between crosses with related and unrelated parents seemed to be useful in combining distance estimates. However, predictions of progeny variance, based on the investigated parental distances, are not reliable enough for practical plant breeding.

Keywords: coefficient of coancestry, genetic distance, *Hordeum vulgare*, major genes, segregation analysis

[3]submitted for publication

Introduction

A plant breeder has to create genetic variation to be able to select new cultivars with combinations of desired traits. This variation is exposed by offspring obtained via crossing of selected combinations of parent genotypes. Selection of parent combinations is usually based on the ability of parents to compensate for their mutual weaknesses. This means that offspring performance is predicted by the average parent value. Besides this parent selection on the basis of predicted mean, it is desirable to select parental combinations producing offspring with a large genetic variation. This variation enables plant breeders to select genotypes that have improved trait values compared to the population average or even compared to the best parent. The larger the offspring variation, the better the opportunities for the breeder to recover transgressive segregants. The magnitude of the offspring variation is determined by the number of segregating genes, i.e., the number of heterozygous loci in the F1, and by their effects. Because the relatedness of the parents is expected to be a good measure of F1 heterozygosity, it has been proposed as a predictor for progeny variation (e.g. Cowen and Frey, 1987a) as well as heterosis (e.g. Smith et al., 1990).

Three sources of relationship information between genotypes are distinguished: (1) geographic information about the origin of the genotypes, (2) pedigree information, and (3) information about plant characteristics (see chapter 2). There are several measures available to quantify this information, e.g. coefficient of coancestry (Malécot, 1948) for pedigree data, also known as kinship coefficient or coefficient of parentage, Euclidean distance (Goodman, 1972), for traits measured on a continuous or ordinal scale, and genetic similarity (Dice, 1945), for binary trait data.

Earlier studies reported poor-to-moderate correlations between parental divergence and progeny variance. Between parental divergence and heterosis a range from zero to moderate correlations are found. Although heterosis and progeny variance are related phenomena, as they both result from F1-heterozygosity, their genetic causes are different. Heterosis is mainly based on dominance effects, while progeny variance is mainly a result of additive effects. Cowen and Frey (1987a) found a positive correlation (r=0.41) between one minus the coefficient of coancestry ($1-f$) and the generalized genetic variance, i.e., the determinant of the trait variance-covariance matrix (Sokal, 1965), for a combination of biomass yield, grain yield and harvest index among F2-derived F3 and F4 lines in oats. They did not find significant positive correlations between this genealogical distance measure and heterosis, although they give quite an extensive overview of earlier research showing positive relationships. Cox and Murphy (1990) observed a moderate correlation between coefficient of coancestry and F2 heterosis in wheat, as well as between a distance measure, based on agronomic and morphological characters, and F2 heterosis. A combination of the two relationship measures predicted F2 yield heterosis better

than the individual measures. In oats, Souza and Sorrells (1991c) found that a combined distance measure, based on pedigree data, qualitative and quantitative morphological characters and biochemical characters, was poorly to moderately correlated with F1 specific combining ability (SCA) for plant height in distant crosses. Coefficients of coancestry showed a moderate correlation with genetic variance for biomass yield among F3-derived F4 lines. Hockett et al. (1993) found indications for increased heterosis for yield in hybrids of unrelated parents compared to hybrids of related parents in barley. This was found within a group of two-row barleys but not in two-row by six-row crosses.

Molecular markers, like RFLP (restriction fragment length polymorphism; Botstein et al. 1980), RAPD (random amplified polymorphic DNA; Williams et al. 1990), and AFLP (Vos et al., 1995), allow the assessment of parental relatedness directly at the DNA level. This was expected to be a great advantage compared to the relationship information previously used. However, in most studies correlations between parental divergence based on molecular markers and progeny variance and/or heterosis are of the same magnitude as mentioned above. Smith et al. (1990) observed a high correlation (r=0.77) between RFLP-based genetic distances and yield heterosis in maize. They used a combination of crosses within and between heterotic parent groups. These two types of crosses were easily distinguished by the genetic distances. Boppenmaier et al. (1993) concluded that RFLP based genetic distances could only predict grain yield heterosis in maize succesfully for crosses within heterotic groups, but not for crosses between groups. Moser and Lee (1994) found moderate correlations between RFLP based genetic distances and genetic variance for plant height and straw yield among F2-derived F3 and F4 lines in oats. They did not find significant correlations between genetic distance and grain yield heterosis. Other relationship measures, based on pedigrees and agronomic traits, did not give better predictions of heterosis and genetic variance. They propose the use of more markers (>68 polymorphisms) and more crosses (>eight) to establish more reliable results. Considering heterosis in wheat, Martin et al. (1995) observed moderate correlations with coefficient of coancestry for kernel weight and protein concentration. Correlations between genetic similarity, based on STS-markers, and heterosis were small and not significant. Investigating four agronomic traits in soybean, Helms et al. (1997) report higher progeny variances among F3-derived F4 lines in three crosses between unrelated parents than in three crosses between related parents. The relatedness was based on pedigree information and could not be confirmed by RAPD-based genetic distance. Manjarrez-Sandoval et al. (1997) found a significant correlation between coefficient of coancestry and variance for yield among soybean SSD-lines, based on data from five crosses. The positive correlation between RFLP-based genetic distance and progeny variance was not significant, which can be attributed to the small number of crosses. Also Burkhamer et al. (1998) report generally positive but non-significant correlations between progeny variance and molecular-marker (STS and AFLP) based genetic distances in wheat. They

observed nine agronomic traits in 12 populations of F3-derived F5 lines. Overall genetic variance appeared to be significantly correlated with coefficient of coancestry and STS-based genetic distance. Also the correlation with a genetic distance measure combining molecular-marker and pedigree information was significant (r=0.64). They conclude, however, that, in general, parental divergence is not a reliable predictor of progeny variance.

In this study we investigate the correlation between progeny variance and a range of parental relationship measures, based on pedigree data, agronomic and morphological characters and AFLP markers, in spring barley. We use a relatively large group of 20 SSD-line populations and establish genetic variance for grain yield in six environments, for thousand kernel weight in three environments, for flowering time in four environments, and for plant height in seven environments. Special attention is paid to the effect of segregating 'major genes' and we will discuss the effect of genotype by environment interaction. Finally we consider the prospects of prediction of progeny variation for practical breeding.

Material and methods

Plant materials

For this study we used 18 two-row spring barley (*Hordeum vulgare* L.) lines, representing parents employed in commercial barley breeding programs in Northwest Europe over the last 20 years. Their pedigree and geographic origin are presented in Table 2.1. These genotypes were used as parents in a partial diallel crossing design (n=18; s=2; Kempthorne and Curnow, 1961) to produce 18 F2 populations (Table 4.1). Reciprocal crosses were made for two parent combinations, increasing the total number of crosses to 20. Single seed descent was performed

Table 4.1. Crosses and their parents. R=reciprocal combination; †='well known' coefficient of coancestry

cross	mother	father	cross	mother	father
1†	Riff	Drossel	11†	Karat	Yriba
2	Baronesse	Forester	12(R2)	Gunhild	GEI-119
3	Baronesse	Bonaire	13†	Gunhild	CEB-9186
4†	Apex	Riff	14	Bonaire	Porthos
5	Porthos	Yriba	15†	CEB-9186	ZE-87-3414
6	Midas	Forester	16†	ZE-87-3414	CEB-9079
7	GEI-119	Midas	17(R2)	GEI-119	Gunhild
8(R1)	Prisma	Apex	18†	Triangel	Georgie
9†	Prisma	Karat	19(R1)	Apex	Prisma
10†	Triangel	Drossel	20†	Georgie	CEB-9079

on 48 F2 plants for each cross until the F5 generation. The F5 plants produced, after one intermediate generation of multiplication, 960 F5-derived F7-lines. These recombinant inbred lines (RILs) were used to obtain estimates of progeny variance for several traits.

AFLP analysis

DNA-extraction followed the CTAB-method described by Van der Beek et al. (1993). AFLP analysis (Vos et al., 1995) followed the protocol described by Van Eck et al. (1995) with modifications by Qi and Lindhout (1997).

Fourteen primer combinations (Table 4.2) generating high numbers of unambiguous polymorphisms in a wide range of barley germplasms were used (Qi and Lindhout, 1997).

Table 4.2. Primer combinations used in AFLP analysis

Number	E+3/M+3 nucleotide extensions	Number	E+3/M+3 nucleotide extensions
E32M50	E+AAC/M+CAT	E35M50	E+ACA/M+CAT
E33M47	E+AAG/M+CAA	E35M54	E+ACA/M+CCT
E33M48	E+AAG/M+CAC	E35M61	E+ACA/M+CTG
E33M50	E+AAG/M+CAT	E38M50	E+ACT/M+CAT
E33M54	E+AAG/M+CCT	E38M59	E+ACT/M+CTA
E33M61	E+AAG/M+CTG	E38M60	E+ACT/M+CTC
E35M47	E+ACA/M+CAA	E38M62	E+ACT/M+CTT

Parent trials

The 18 parent lines plus two additional standards (Magda and Vada) were tested at several locations in the years 1993 to 1996 (Table 4.3). In 1995 and 1996 they were used as standards in the recombinant inbred line trials, described in the next paragraph. In 1994 the parents and Magda and Vada were used as standards in a small plot yield trial of F4-material. In this trial all standards occurred six times per location. The remaining trials had two replicates and we applied a partially balanced incomplete block design with four plots per incomplete block. Plot sizes and sowing dates are presented in Table 4.3. Observed traits were plant height (cm), flowering time (°C.days), thousand kernel weight (g) and grain yield (kg/ha), although not all traits were observed in all trials (Table 4.3). Flowering time was defined as the temperature sum from emergence to decimal stage 49 (Zadoks et al. ,1974). All trials were kept free from diseases.

RIL-trials

The 960 recombinant inbred lines (RILs) were tested at several locations in 1995 and 1996 (Table 4.3). In 1996 three trials (Lelystad: 96-3, Rilland: 96-4, and Ottersum: 96-5) each

Table 4.3. Field trial description. p=present in trial; †=only part of RILs present in trial; h=plant height; f=flowering time; k=thousand kernel weight; y=grain yield

trial	location	soil	year	plot size (m^2) (width × length (m))	sowing date	parents	RILs	observed traits
93-1	Wageningen	clay	1993	5.55 (1.5 × 3.7)	2 April	p		hky
93-6	Wageningen	sand	1993	5.55 (1.5 × 3.7)	26 March	p		hky
94-1a	Lelystad	clay	1994	0.94 (0.625 × 1.5)	31 May	p		y
94-2a	Wageningen	clay	1994	0.94 (0.625 × 1.5)	1 June	p		y
94-1b	Lelystad	clay	1994	8.55 (1.5 × 5.7)	22 April	p		hky
94-2b	Wageningen	clay	1994	8.55 (1.5 × 5.7)	28 April	p		hfky
94-6	Wageningen	sand	1994	8.55 (1.5 × 5.7)	27 April	p		hfky
95-1	Swifterbant	clay	1995	9.0 (1.5 × 6.0)	15 May	p	p	hfy
95-2	Wageningen	clay	1995	9.0 (1.5 × 6.0)	11 May	p	p	hfky
95-6	Wageningen	sand	1996	9.0 (1.5 × 6.0)	20 April	p		hfy
96-1	Swifterbant	clay	1996	9.0 (1.5 × 6.0)	1 April	p	p	hfk
96-2	Wageningen	clay	1996	9.0 (1.5 × 6.0)	19 March	p	p	hfky
96-3	Lelystad	clay	1996	5.32 (1.4 × 3.8)	18 April	p	p†	hy
96-4	Rilland	clay	1996	3.6 (1.5 × 2.4)	19 March	p	p†	hy
96-5	Ottersum	sand	1996	4.65 (1.5 × 3.1)	18 March	p	p†	hy
96-6	Wageningen	sand	1996	9.0 (1.5 × 6.0)	26 March	p		y

contained one third of the lines (i.e., 16 lines) of each cross. They are called 'partial' locations, while the other trials are indicated as 'complete' locations. The 18 parent lines and cultivars Magda and Vada were added as standards. Each trial included two replicates. Each standard occurred six times per replicate at the 'complete' locations and two times per replicate at the 'partial' locations. All genotypes were randomised according to a partially balanced incomplete block design with 8 plots per block. The constraint that two genotypes do not occur more than once together in the same block, extended over all RIL trials. Sowing dates and plot sizes are presented in Table 4.3.

Observed traits were grain yield (kg/ha), thousand kernel weight (g; 0% moisture), plant height (cm) and flowering time (°C.days). Grain dry matter content at harvest was measured on a 100g-sample. Lodging was scored on a scale from 0 (no lodging) to 5 (severe lodging) around decimal stage 69 (Zadoks et al., 1974), except for trial 95-2. In this trial lodging was observed at stage 83 as there was hardly any lodging in earlier stages.

Statistical analysis

Parent trials

Per parent trial an analysis of variance was performed for each trait using a linear mixed model with fixed effects for the 18 parents and 2 additional standards. Fixed replicate effects and random incomplete block effects were included in the model whenever they appeared significant (α=0.05). Average plant height of the two adjacent plots was used as a covariate in the analysis

of the yield data. Parent least squares means were calculated for use in agronomic distance estimation.

RIL-trials

RIL-trial data were analysed per trait (yield, thousand kernel weight, plant height, flowering time and lodging) using average information REML (Gilmour et al., 1995). Analysis of the plot data was performed per year by location combination (environment), because an overall analysis appeared not feasible due to computational limitations. The linear mixed model included fixed effects for standards, crosses and strips of adjacent incomplete blocks. The block effects were assumed random, as well as the line within cross effects. In the analysis of residuals and whenever a hypothesis considered the specific lines that were present in the trial, the line effects were assumed to be fixed. For the analysis of lodging data we fitted a proportional odds model (McCullagh & Nelder, 1989). Residual analysis was performed to trace outliers among the data. These observations were excluded from the final analyses.

To overcome computational limitations we performed a combined analysis over years and locations (environments) by using the least squares means for the lines as input data for an analysis of variance. The linear mixed model included fixed effects for standards, crosses, standard by environment interaction and cross by environment interaction. A random line effect over environments was included to obtain an estimated line variance V per cross for every trait.

Generalised genetic variances (GGV) per cross were calculated as the determinant of the $n \times n$ variance-covariance matrix for the traits (Sokal, 1965), where n is the number of trait by environment combinations. The determinant is the product of the n eigenvalues of the matrix. To make the GGVs comparable to the Vs, the GGVs were transformed by taking their n^{th} root. The overall GGV was based on the available least squares means of four traits (plant height, flowering time, thousand kernel weight, grain yield) from the 4 'complete' RIL trials (95-1,2 and 96-1,2). We calculated trial specific GGVs on the basis of the available trait data from each single 'complete' RIL trial. We also calculated trait specific GGVs on the basis of single trait data from all RIL trials. In contrast to the line variances V over environments, the trait specific GGVs include variation caused by genotype by environment interaction. For plant height and for grain yield, these GGVs were geometric means of three 'partial' GGVs. These 'partial' GGVs were calculated from data from a subset of 16 RILs per cross, tested in one of the 'partial' trials (96-3, 96-4, 96-5). We used trait data from the partial environment together with data from the 'complete' environments.

Segregation analysis

On the basis of parent information we expected several crosses to segregate for ‘major genes’, e.g. the *denso* gene (Haahr & Von Wettstein, 1976) and the *ert-g* gene (Thomas et al., 1984). This segregation increases the variance among the offspring lines. To investigate the effect of segregating ‘major genes’ on the relationship between parental divergence and progeny variance, segregation analysis was performed to trace these genes. Therefore we applied a robust mixture model (McLachlan & Basford, 1988) on the trait least squares means of the lines (plant height, flowering time, thousand kernel weight or yield) per cross. The model was fitted using the idea for an iterative EM-algorithm described by Jansen (1993). The algorithm was implemented in SAS-IML (SAS Institute Inc., 1989). The presence of a segregating ‘major gene’ was accepted on the basis of a likelihood ratio test (α=0.05), in which the unimodal model (H_0: no ‘major gene’ segregating) was tested against the bimodal model (H_1: ‘major gene’ segregating). Confidence thresholds for the likelihood ratio were obtained by Monte Carlo sampling from a multivariate normal distribution with a variance-covariance matrix based on the observed data. When a bimodal model was accepted, we refer to this as a 'major gene' segregating in the cross. Once a ‘major gene’ was accepted, this factor was used as a covariable in the mixture model, while looking for additional ‘major genes’ . This procedure was repeated until no further ‘major genes’ were detected.

With the use of the postulated ‘major genes’, genetic variance between lines could then be divided into two parts: variance caused by the hypothesised ‘major genes’ and variance resulting from the segregation of other -‘minor’- genes. The latter variance is therefore called ‘minor gene’ variance (mgV). It is estimated by using the segregating ‘major genes’ and their interactions as explanatory variables in the analysis of variance. In the analysis over environments we included the effect of ‘major gene’ by environment interaction.

‘Minor gene’ generalised genetic variances (mgGGV) were calculated similarly as GGVs. They are based on least squares means corrected for the effects of ‘major genes’ and mutual interactions of ‘major genes’.

Genetic-distance estimation

Markers were scored following the procedure described in chapter 2 in which redundant AFLP markers within primer combinations were discarded. Genetic similarities (*gs*) were calculated following Nei and Li (1979): $gs_{ij}=\frac{2N_{ij}}{N_i+N_j}$,where N_{ij} is the number of bands present in both genotypes *i* and *j*, N_i is the number of bands present in genotype *i* and N_j is the number of bands present in genotype *j*. In the case of a missing observation for a marker in genotype *i* and/or *j*, this marker was not included in the calculation of gs_{ij}. The accuracy of *gs*-estimates as influenced by sampling and missing marker data was assessed by taking bootstrap samples

(Efron and Tibshirani 1993) from all markers, including polymorphic as well as monomorphic markers. Bootstrap standard-deviation estimates were based on 1000 samples.

On the basis of pedigree data the coefficient of coancestry f (Malécot, 1948) was calculated for the parental combinations in Table 4.1, following the assumptions of Van Hintum and Haalman (1994). Only 10 combinations had a 'well known' f (Table 4.1), as defined in chapter 2, and were used in further analysis.

AFLP-based genetic distance was calculated as $1-gs$. Analogously the coefficient of coancestry was converted into a genetic distance measure $1-f$.

Morphological distances (*md*) were calculated on the basis of 25 morphological traits described in chapter 2. The observed parent data were standardised and a principal components analysis was performed. Principal components with an eigenvalue greater than an arbitrary value K=1.0, were used to calculate generalised distances between the parents (Goodman, 1972).

A second multivariate distance between the 18 parent genotypes and the two standards (Magda and Vada) was calculated on the basis of four agronomic characters: plant height (cm), flowering time (°C.days), thousand kernel weight (g) and grain yield (kg/ha). We used the least squares means for the traits from the different environments as separate variates in the analysis, so the agronomic distance (*agd*) was based on 43 trait by environment combinations (Table 4.3). The *agd* was calculated following the procedure of Goodman (1972), as used for the calculation of morphological distance.

Finally the four distance measures ($1-gs$, $1-f$, *agd*, *md*) were combined in two ways. For both methods we standardised the distances for each parental distance matrix to make them comparable. The pooled distance measure *pd1* is the first principal component from a principal components analysis of the standardised distances. The pooled distance *pd2* is based on the distinction between relatively close parent combinations and relatively distant parent combinations. Burstin and Charcosset (1997) point out that small morphological and agronomic distances between two genotypes do not necessarily mean that these genotypes are closely related. Therefore these two distance measures are only used for the calculation of *pd2* in unrelated parent combinations. The other two distance measures, based on molecular markers and pedigrees, reliably predict the distance between closely related genotypes. For unrelated parent combinations differences between molecular markers may not be representative for the QTL heterozygosity causing variation in the offspring (Charcosset et al., 1991). The assumption of equally unrelated ancestors and the assumption of each parent contributing exactly 50% to the offspring genotype cause large biases in the calculated coefficients of coancestry, especially in unrelated genotype combinations (see chapter 2). So the average of the standardised $1-gs$ and

Table 4.4. Method of calculating the pooled distance *pd2* on the basis of the standardised agronomic (agd_{st}), morphological (md_{st}) and AFLP-based genetic distance ($(1-gs)_{st}$) and the standardised coefficient of coancestry ($(1-f)_{st}$), where $pd2_{close}=½((1\text{-}gs)_{st}+(1\text{-}f)_{st})$ and $pd2_{distant}=½(agd_{st}+md_{st})$

		if $pd2_{distant}$ (for unrelated parent combinations)		
		≤0	>0	
if $pd2_{close}$ (for related parent combinations)	≤0	$pd2_{close}$	$pd2_{close}$	if $(pd2_{close}+pd2_{distant})\leq 0$
			$pd2_{distant}$	if $(pd2_{close}+pd2_{distant})>0$
	>0	$(pd2_{close}+pd2_{distant})/2$	$pd2_{distant}$	

$1-f$ serves as distance measure $pd2_{close}$ for related parents and the average of the standardised *agd* and *md* serves as distance measure $pd2_{distant}$ for unrelated parents. The degree of relatedness of two genotypes was decided upon the values of these two averages and the pooled distance *pd2* was calculated following the decision rules in Table 4.4. Two additional pooled distances *pd1*' and *pd2*' were calculated without using the coefficient of coancestry. So $pd2_{close}$ was equal to the standardised value of $1-gs$. In this way the 10 crosses without 'well known' coefficients of coancestry could also be used in the analysis.

Table 4.5. RIL trials characterised by average trait values, observed over 20 populations and 20 standards and root mean square errors (root mse) obtained by variance analysis. tkw = thousand kernel weight.

trial	mean					root mse			
	plant length (cm)	flowering time (°C.days)	lodg-ing (0-5)	tkw (g)	yield (kg/ha)	plant length (cm)	flowering time (°C.days)	tkw (g)	yield (kg/ha)
95-1	81	618	0.6		5452	3.1	9.7		229
95-2	77	593	1.5	46.4	6048	3.1	11.4	1.5	220
96-1	90	690	3.1	44.2		3.4	9.6	2.2	
96-2	80	643	2.7	48.0	9127	3.3	8.6	1.6	293
96-3	94		1.5		7404	2.8			301
96-4	87		2.8		9813	3.3			453
96-5	89		0.7		9067	2.9			261

Results

RIL trials

Average trait values and root mean square errors per trial are presented in Table 4.5. Square roots of the estimated between line variances V per cross over environments are presented in Table 4.6, as well as the n^{th} root of the generalised genetic variance over all available traits from the 'complete' trials. Due to severe hail storm damage, we did not obtain yield data from Swifterbant. We could, however, sample spikes from this location to observe thousand kernel weights.

Table 4.6. Square root of variance and 'minor gene' variance for plant height, flowering time, thousand kernel weight (tkw), and yield, per cross, analysed over RIL trials (95-1,2, 96-1,..,5), n^{th} root of generalised genetic variance (GGV) and 'minor gene' generalised genetic variance (mgGGV) over all trait by 'complete' environment (95-1,2 and 96-1,2) combinations and CVs for all variances.

	√(variance)				$(GGV)^{1/n}$	√('minor gene' variance)				$(mgGGV)^{1/n}$
cross	plant height (cm)	flowering time (°C.days)	tkw (g)	yield (kg/ha)		plant height (cm)	flowering time (°C.days)	tkw (g)	yield (kg/ha)	
1	7.41	28.3	3.67	397	12.2	2.47	9.6	2.83	322	10.1
2	7.25	23.5	2.68	292	11.6	2.91	4.5	0.98	192	7.2
3	4.92	23.7	2.78	252	11.1	2.58	12.4	2.11	246	10.0
4	8.06	29.9	1.63	254	10.9	2.71	8.7	0	173	9.1
5	3.19	11.7	1.87	197	10.4	3.19	11.7	1.87	197	10.4
6	9.11	47.8	4.09	461	13.6	1.45	29.8	2.18	268	9.3
7	3.09	41.0	1.86	232	11.1	1.80	7.4	1.40	190	8.3
8	7.58	24.9	1.93	357	12.5	4.04	10.0	1.97	237	10.6
9	2.61	24.0	1.27	147	11.4	2.61	24.0	1.27	147	11.4
10	2.91	23.2	2.05	199	11.5	2.71	10.6	1.22	182	10.2
11	8.13	42.6	1.97	393	14.0	2.67	33.6	2.05	177	12.2
12	7.36	37.3	2.85	153	13.2	3.36	17.3	1.49	132	10.1
13	8.43	39.6	2.66	234	12.5	3.17	17.3	2.36	190	10.7
14	5.32	29.9	1.95	356	12.4	3.60	8.2	1.11	297	9.0
15	6.76	46.8	1.82	289	13.1	4.60	29.9	1.25	260	10.1
16	4.99	37.7	2.10	336	12.7	3.57	17.0	2.08	267	11.8
17	6.97	37.8	2.99	217	13.5	4.38	14.6	1.72	173	10.7
18	2.14	21.2	2.01	241	10.6	1.85	13.4	1.97	236	9.0
19	6.95	27.5	2.08	344	12.5	3.64	12.2	2.02	179	10.9
20	6.55	27.7	2.28	492	11.1	3.23	14.6	2.35	254	9.6
CV	0.60	0.57	0.64	0.65	0.17	0.52	1.07	0.59	0.47	0.23

Table 4.7. Number of postulated segregating 'major genes' per cross. Known 'major genes' that are segregating, are specified: *denso* (Haahr & Von Wettstein, 1976) and *ert-g* (Thomas et al., 1984).

cross	number of 'major genes'	known 'major genes' segregating
1	2	*denso*
2	4	*denso*
3	1	
4	2	*denso*
5	0	
6	3	*denso, ert-g*
7	3	
8	2	*denso*
9	0	
10	2	
11	2	*denso*
12	3	*ert-g*
13	2	
14	4	
15	3	
16	1	*denso*
17	3	*ert-g*
18	2	
19	2	*denso*
20	2	*denso*

Segregation analysis

As a result of segregation analysis we were able to postulate 0 to 4 segregating 'major genes' per cross (Table 4.7). Whenever we expected segregation of the *denso* or the *ert-g* gene in a cross, on the basis of the pedigrees of the parents, we were able to identify this segregation by visual inspection of the distribution as well as by using the mixture model. The postulated 'major genes' and their mutual interactions explained part of the trait variance V over environments. The average explained proportions were: 47% for plant height, 55% for flowering time, 27% for thousand kernel weight, and 25% for yield (Table 4.6). In a few cases 'minor gene' variances that were slightly larger than V, were found. They result from crosses in which 'major genes' predominantly explain line by environment interaction variance. As a result of the separation of this interaction, the average line effect over environments can be estimated more accurately, leading to an increased between line variance over environments.

Genetic-distance estimation

In total 1248 markers were used to estimate genetic similarities (*gs*) and 36 % of them showed polymorphism among the 18 parent genotypes. Genetic distances 1 – *gs* for all possible parent combinations ranged from 0.031 to 0.085 with an average of 0.064. The bootstrap standard deviations for these (1 – *gs*) values ranged from 0.0041 to 0.0065. The genetic distances for the 20 crosses under investigation ranged from 0.044 to 0.075 with an average of 0.064.

'Well known' coefficients of coancestry (*f*) were obtained from the pedigree data. The parental divergence 1 – *f* ranged from 0.436 to 0.950 with an average of 0.845. The 10 crosses in this study with reliable pedigree data for both parents (Table 4.1) had a value of 1 – *f* ranging from 0.436 to 0.904 with an average of 0.798.

Morphological distances (*md*) among the 18 parents, based on 10 principal components, ranged from 1.88 to 5.86 with an average of 4.32 (see chapter 2). The *md* values of the 20 investigated crosses ranged from 2.06 to 5.46 with an average of 4.14.

Agronomic distances (*agd)*, calculated over all trait by environment combinations mentioned in Table 4.3, were based on 7 principal components explaining 90% of the variation. The values for *agd* between the 18 parent lines ranged from 1.42 to 5.89 with an average of 3.67. The sample of 20 crosses had *agd* values ranging from 2.33 to 4.87 with an average of 3.72.

The pooled distances among the 18 parents ranged from – 3.30 to 2.62 for *pd1*, and from – 3.20 to 1.48 for *pd2*. The principal component *pd1* explained 38% of the variation of the four distance measures. For the 10 crosses with 'well known' *f* values the ranges were [– 2.91, 1.43] for *pd1* and [– 3.20, 1.13] for *pd2* and the averages were – 0.12 for *pd1* and – 0.07 for *pd2*. Pooled distances based on only 1 – *gs*, *md* and *agd* ranged from – 4.27 to 2.61 for *pd1*', and from – 3.89 to 2.06 for *pd2*'. Principal component *pd1*' explained 50% of the variation of the three distance measures. For the 20 crosses in Table 4.1 *pd1*' ranged from – 2.35 to 2.11 with an average of – 0.10 and *pd2*' ranged from – 2.39 to 1.22 with an average of – 0.17.

Relationship between genetic distance and progeny variation

Correlation coefficients between the parental divergence measures and the estimated offspring variation are presented in Table 4.8. A bootstrap procedure (Efron and Tibshirani, 1993), taking 10,000 samples from the set of 20 parent combinations, was used to test whether these correlation coefficients were significantly (α=0.05) higher than zero. Correlation coefficients were mainly positive but small and non-significant. AFLP-based genetic distance (1 – *gs*) and pooled distances showed generally the 'best' correlations with progeny variance. Agronomic distances, based on parent data for a single trait or environment, showed several significant

Table 4.8. Correlation coefficients between parental divergence measures and progeny variances. 1−*gs*=AFLP-based genetic distance; 1−*f*=pedigree-based genetic distance; *md*=morphological distance; *agd*= agronomic distance; *pd1*, *pd2*=pooled distances based on 1−*gs*, 1−*f*, *md* and *agd*; *pd1*', *pd2*'=pooled distances based on 1−*gs*, *md* and *agd*; *agd*(tr/env)=agronomic distance based on data for one trait from all environments or on data for all traits from one environment, the single trait or the single environment is given in the second or third column; V/mgV=variance/'minor gene' variance analysed over environments; GGV/mgGGV=generalised genetic variance/generalised 'minor gene' variance, based on specified set of *n* trait by environment combinations; V'/mgV' =variance/'minor gene' variance analysed per environment and correlation coefficients averaged over environments; h=plant height; f=flowering time; k=thousand kernel weight; y=yield; A=95-1; B=95-2; C=96-1; D=96-2; E=96-3; F=96-4; G=96-5; *=0.01<P<0.05; **=P<0.01

variance	traits	environments	1−*gs*	1−*f*	*md*	*agd*	*pd1*	*pd2*	*pd1*'	*pd2*'	*agd* (tr/env)
V	h	ABCDEFG	0.31	-0.24	0.17	0.23	-0.24	-0.23	0.31	0.31	0.37*
mgV	h	ABCDEFG	0.43*	0.18	-0.04	0.28	0.42	0.37	0.26	0.38*	0.41
V	f	ABCD	0.03	0.07	0.22	0.24	0.31	0.28	0.23	0.08	0.41*
mgV	f	ABCD	0.27	0.24	0.26	-0.12	0.39	0.29	0.20	0.27	0.41*
V	k	BCD	0.09	0.27	0.16	0.24	0.25	0.36	0.22	0.23	0.29
mgV	k	BCD	0.02	0.47	-0.23	0.06	0.22	0.42	-0.09	0.16	-0.07
V	y	ABDEFG	0.02	0.13	-0.03	0.00	0.01	0.15	0.00	0.12	0.16
mgV	y	ABDEFG	-0.29	0.23	0.04	0.05	0.33	0.46*	-0.07	0.09	0.08
$GGV^{1/n}$	hfky	ABCDEFG	0.51**	0.19	0.29	0.27	0.53*	0.52*	0.47**	0.47*	
$mgGGV^{1/n}$	hfky	ABCDEFG	0.47	0.17	-0.03	-0.04	0.47	0.51	0.16	0.40	
$GGV^{1/n}$	h	ABCDEFG	0.48*	-0.26	0.25	0.35*	0.13	0.10	0.46**	0.51**	0.58**
$mgGGV^{1/n}$	h	ABCDEFG	0.39	-0.17	-0.10	0.09	0.39	0.33	0.14	0.35	0.24
$GGV^{1/n}$	f	ABCD	0.27	0.14	0.36	0.18	0.49*	0.45*	0.37	0.31	0.34
$mgGGV^{1/n}$	f	ABCD	0.31	0.15	0.20	-0.15	0.42	0.37	0.16	0.29	0.22
$GGV^{1/n}$	k	BCD	0.48*	0.48	0.14	0.40**	0.55	0.67*	0.43**	0.47**	0.13
$mgGGV^{1/n}$	k	BCD	0.26	0.59	-0.42	-0.02	0.30	0.50	-0.13	0.10	-0.08
$GGV^{1/n}$	y	ABDEFG	0.13	0.22	0.16	0.20	0.22	0.30	0.22	0.30	0.17
$mgGGV^{1/n}$	y	ABDEFG	0.34	0.34	-0.07	-0.16	0.37	0.49	0.03	0.40	-0.24
$GGV^{1/n}$	hfy	A	0.22	0.20	0.13	0.12	0.20	0.19	0.20	0.21	0.29
$mgGGV^{1/n}$	hfy	A	0.41	0.41	-0.02	-0.13	0.41	0.40	0.10	0.36	0.06
$GGV^{1/n}$	hfky	B	0.24	0.31	0.14	0.18	0.37	0.37	0.25	0.24	0.52**
$mgGGV^{1/n}$	hfky	B	0.36	0.29	-0.14	-0.18	0.30	0.31	0.00	0.22	0.23
$GGV^{1/n}$	hfk	C	0.25	0.27	0.12	0.17	0.25	0.36	0.23	0.28	0.43**
$mgGGV^{1/n}$	hfk	C	0.30	0.39	-0.15	-0.07	0.41	0.57	0.01	0.29	0.16
$GGV^{1/n}$	hfky	D	0.29	0.19	0.18	0.23	0.25	0.31	0.31	0.34	0.28
$mgGGV^{1/n}$	hfky	D	0.48*	0.36	0.04	0.10	0.52	0.61*	0.25	0.53*	-0.28
V'	h	ABCDEFG	0.20	-0.22	0.04	0.13	-0.22	-0.21	0.16	0.19	0.26*
mgV'	h	ABCDEFG	0.13	0.07	-0.24	0.08	0.21	0.15	-0.05	0.09	0.08
V'	f	ABCD	0.02	0.05	0.22	0.22	0.32	0.30	0.22	0.10	0.37*
mgV'	f	ABCD	0.25*	0.11	0.22	0.00	0.34	0.33	0.22	0.29*	0.21
V'	k	BCD	0.14	0.29	0.16	0.25*	0.31	0.41	0.24*	0.26	0.27
mgV'	k	BCD	0.04	0.42	-0.23	0.06	0.20	0.38	-0.08	0.15	-0.07
V'	y	ABDEFG	0.00	0.02	0.15	0.13	0.14	0.23	0.13	0.14	0.20
mgV'	y	ABDEFG	-0.05	0.21	0.05	0.03	0.31	0.38*	0.02	0.12	0.00

correlations with progeny variances for the same trait or environment. However, elimination of 'major gene' effects often resulted in a clear decrease of the correlation-coefficient values.

Discussion

Correlations between parental divergence measures were generally poor, as observed in chapter 2, for *gs*, *f* and *md*. Correlation coefficients with *agd* were 0.083 (1–*gs*), 0.180 (1–*f*; only 'well known' *f*), and 0.372 (*md*). This lack of correspondence among distance measures is confirmed by the relatively small proportions of variation explained by their first principal component *pd1* or *pd1*'. This is mainly a result of inaccuracies and incorrect assumptions in the calculation of the different genetic distance measures. Most of these imperfection were described by Burkhamer et al. (1998) and in chapter 2. Further, Burstin and Charcosset (1997) conclude that small agronomic distances do not necessarily indicate relatedness. We may add that large agronomic trait differences between parents are sometimes due to an allelic difference for only one 'major gene'. Therefore large *agd* values are not necessarily indicating large parental divergences.

Especially for yield we observed a lack of association among progeny variances for the different environments. The average rank correlation coefficient between RIL variances obtained from different environments was 0.67 for plant height, 0.72 for flowering time, 0.70 for thousand kernel weight, and only 0.29 for grain yield. The average rank correlation between 'minor gene' variances was smaller: 0.27 for plant height, 0.55 for flowering time, 0.54 for thousand kernel weight, and 0.10 for grain yield. Inaccurracy of yield variance estimates is not a likely cause of this heterogeneity of cross variances, because the standard errors for the variance components of the different traits are similar in size: about 30% of the variance value and about 40% of the 'minor gene' variance value. The variation in yield variance among crosses is similar to that of other traits: CVs are comparable (Table 4.6). Therefore we assume that between RIL variance for yield in different environments is a result of different genomic distributions of segregating yield genes and their effects. This means that yield is controlled by genes with large and small effects, even after the elimination of the postulated 'major genes'. In our study we observed that 'major gene' effects for yield differed over environments. The assumption of QTL by environment interaction for yield is also corroborated by QTL mapping studies in barley, e.g. Hayes et al. (1993) and Tinker et al. (1995). Due to this interaction, it is likely that only part of the yield 'major genes' have been found in the segregation analysis. Most of the 'major genes' were postulated on the basis of segregation patterns for plant height, lodging and flowering time.

The correlation between parental divergence and progeny variance is generally poor (Table 4.8). This lack of association, as far as it is not caused by inaccuracies in the genetic

distance and variance estimates, is basically explained by Charcosset et al. (1991). They state that RIL variance is caused by a limited number of segregating QTL. Part of the QTL positions at the genome are not covered by genetic distance data and, conversely, part of the genetic distance data are based on genome regions without QTL. The poor association of parental distance measures already shows that distance information from different sources depends on the monitored parts of the genome. Also the different QTL, found in QTL studies as mentioned above, and responsible for RIL variance for different traits in different environments, make clear that reliable variance prediction based on a single genetic distance measure is virtually impossible as distance information never matches with all sets of QTL.

The association of genetic distance measures with trait variance at individual trials (V') was generally absent. Agronomic distance (*agd*) based on plant height differences among the parents showed significant correlation with plant height variance. Analogously, *agd* based on parental flowering time data was significantly correlated with RIL variance for flowering time. These correlations are also present for variances V estimated over environments and single trait generalised genetic variances (GGV). For height and flowering time agronomic distances apparently contain parental divergence information from the DNA positions of the most important QTL. Elimination of this 'major gene' effect in the variances usually results in a decrease of the correlation coefficients. For the generalised genetic variances, based on data from a single environment, we observe a similar pattern. Agronomic distances based on parental differences in the same environment significantly correlate with GGV, but not with mgGGV.

The AFLP-based genetic distance seems to have higher correlation coefficients with 'minor gene' variances than with variances including 'major gene' effects. However, this is not always the case. Apparently, $1-gs$ is also predicting the absence or presence of segregating 'major genes' for plant height and thousand kernel weight, as it is significantly correlated with generalised genetic variances for these traits. The significant correlations of $1-gs$ with 'minor gene' variance is explained by the fact that 'minor genes' as well as AFLPs are assumed to have positions dispersed all over the genome. Also the association of $1-gs$ and generalised genetic variance GGV can be attributed to this fact. Presumably the number of QTL responsible for the value of GGV is that high that QTL are positioned basically everywhere on the genome. This association is confirmed by results of Burkhamer et al. (1998).

The number of crosses with 'well known' pedigrees is too small to establish significant correlations between $1-f$ and progeny variance. However, especially pooled distance measure *pd2* is significantly correlated with several RIL variances, including 'minor gene' variance for yield. We also observe several significant correlations of *pd1*' and *pd2*' with genetic variance. This confirms conclusions of Cox and Murphy (1990) and Souza and Sorrells (1991c) that combining distance measures results in better predictions of progeny variance. Observing the correlation coefficients in Table 4.8, it seems that a distinction between related and unrelated

crosses, as used in the calculation of *pd2* and *pd2'*, performs slightly better than a rather straightforward principal components analysis (*pd1* and *pd1'*) to combine genetic distance measures.

The segregation of 'major genes', e.g. *denso* and *ert-g*, is usually predicted on the basis of the parent pedigree information and the desired allele can be selected visually in early generations and/or in small plots. In practical breeding programmes often both parent genotypes contain the desired alleles for these genes. In both cases only 'minor gene' variation can be subsequently exploited in SSD or DH populations. If a breeder wants to predict 'minor gene' variation, AFLP-based genetic distance or a genetic distance combining several sources of distance information perform relatively well. However, the correlation coefficients between parental relatedness and RIL variance do not seem to be high enough to be really useful in practical breeding.

Conclusion

Genetic distance measures based on different sources of parent information do not correspond well. This is the result of inaccuracies in the estimation as well as differences in the representation of genomic divergence between the parents. Segregation analysis of the RIL populations resulted in 43 postulated 'major genes' with important contributions to the genetic variance of the investigated traits. Agronomic distances between the parents based on plant height, flowering time, or thousand kernel weight predict this 'major gene' variation well. AFLP-based genetic distance show poor, but significant correlations with some 'minor gene' variances and generalised genetic variance over all traits. The distance as well as the variances may be related to the overall genomic divergence. However, QTL not linked to AFLP markers and AFLP markers not linked to QTL result in poor observed correlations (Charcosset et al., 1991). Lack of reliable pedigree information for half of the crosses prohibits a clear statement about the predictive value of the coefficient of coancestry. Pooled distance measures seem to perform equally well as AFLP-based genetic distance. Among all distance measures they show the strongest correlation with progeny variance for yield. However, in general progeny variance predictions on the basis of the investigated genetic distance measures are not reliable enough for practical breeding.

Acknowledgements

The authors thank Cebeco Zaden B.V., Vanderhave Research and Zelder B.V. for making available some of their advanced breeding lines and extensive trial facilities. They thank CPRO-DLO and CIMMYT, Mexico for the use of trial facilities. They also thank Piet Stam, Johan Dourleijn, Ritsert Jansen, Johannes Kalma, and Derk Gesink for their contributions to this study.

5

Prediction of progeny variation in barley crosses using parental relationship measures. II. Measures based on genetic map information and marker-trait associations among parents [4]

Johan W. Schut and Piet Stam

Abstract

Progeny variation for several agronomic traits in 20 crosses of spring barley was predicted by two AFLP-based genetic distances. The first genetic distance, *mgd*, was estimated by the use of genetic map information. Markers were weighted for the marker density at their map position in order to obtain a more uniform representation of the genome. In comparison with the correlation coefficient between unweighted genetic distance and progeny variance, the use of *mgd* did not result in higher correlation coefficients. The second genetic distance, *sgd*, was based on selected markers that showed significant association with the investigated trait in the parent population. Several marker selection procedures were compared and the highest correlations between *sgd* and progeny variance were found when an additional randomly chosen marker was included in the model for marker-trait association. An average P-value for the F-test of the marker under investigation was obtained by using all other markers, one by one, as random marker in the model. The optimum selection threshold for the P-value was 0.005. In the case of flowering time, an extension of the model with marker by environment interaction appeared useful. In comparison with the correlation between genetic distance based on all markers and progeny variance for the different traits, the use of marker selection resulted in higher correlation coefficients. However, sometimes the correlation coefficients were high mainly as a result of segregating 'major genes'. We conclude, that correlations between *sgd* and progeny variance are not high enough to be useful in practical breeding.

Keywords: AFLP, genetic distance, genetic map, *Hordeum vulgare*, marker selection

Introduction

[4]submitted for publication

Prediction of progeny performance on the basis of parent information is very important for practical plant breeding. If one is able to select the right parental combination, i.e., the combination producing a progeny that includes a genotype which is better than the existing cultivars, much time, space, and effort can be saved. Prediction of offspring mean is usually based on the midparent average for the trait (Bos and Caligari, 1995). Prediction of offspring variance causes more difficulties. This genetic variance is related to the number of segregating genes and their effects. The number of heterozygous loci in the F1 that will segregate in following generations, can be predicted on the basis of the relatedness of the parents. So, parental divergence could indirectly serve as a predictor for progeny variance. Many authors have tried to find empirical evidence for this relationship by using relationship measures based on parental pedigree data, morphological and agronomical characters, and biochemical and molecular marker information. A brief overview is given in chapter 4. The general conclusion of these investigations is that the correlation between parental divergence and genetic variance within a cross is positive, but weak.

Many reasons have been suggested why parental relationship measures are poor predictors. The use of pedigree data is hampered by unrealistic assumptions that have to be made for the calculation of a coefficient of coancestry (Cowen and Frey, 1987a; Burkhamer et al., 1998). Estimation of progeny variances is not very accurate and is subject to large sampling errors (Burkhamer et al., 1998). In some cases the number of crosses is considered too small to establish a significant relationship between parental divergence and offspring variation (Moser and Lee, 1994; Manjarrez-Sandoval et al., 1997). This relationship is also affected by the parent population structure, i.e., the presence of one or more genetically distinct groups among the parents (Souza and Sorrells, 1991c; Boppenmaier et al., 1993; Charcosset and Essioux, 1994). The use of quantitative morphological traits to estimate genetic distance is also of limited value as these traits tend to be strongly affected by only a few ‘major gene’ loci (Souza and Sorrells, 1991c) and do therefore not present information about the whole genome. Burstin and Charcosset (1997) show that there exists no straightforward linear relationship between phenotypic differences and genotypic differences. In the case of genetic distances based on biochemical or molecular markers, representation of the genome can be poor, depending on the type of markers (Powell et al., 1996) and the number of markers (Moser and Lee, 1994). Finally, it is suggested that genetic variances for agronomic traits like yield are based on the segregation of a limited number of genes with large effects (Souza and Sorrells, 1991c; Moser and Lee, 1994; see also chapter 4).

Several authors have suggested ways to improve the prediction of F1-heterozygosity of crosses and/or its resulting breeding behaviour. This can be heterosis as well as progeny variance, although, as mentioned in chapter 4, these phenomena are based on different effects of the heterozygous loci. Charcosset et al. (1991) showed in a theoretical study that the

correlation between heterosis and heterozygosity at molecular marker loci is decreased by QTL that are not marked by marker loci and by marker loci that do not mark QTL. Empirical studies testing improvements for heterosis prediction show varying results. Dudley et al. (1991) used a genetic distance based on 29 RFLP loci that were significantly associated with maize hybrid yield, to predict specific combining ability (SCA) for parent combinations. Correlation was poor and even smaller than correlation between SCA and genetic distance based on all RFLPs. However, Zhang et al. (1994) found a strong correlation (r=0.77) between F1 heterosis for yield in rice and genetic distance based on 16 RFLP and SSR markers that were significantly associated with F1 yield. Barbosa-Neto et al. (1996) could not show a significant relationship between heterosis for yield in wheat and genetic distance based on RFLP markers associated with parent yields. In this case marker-trait associations and predicted values were not based on the observed hybrids. One could, however, argue that a marker-trait association found in an arbitrary sample of parent lines does not necessarily reflect real linkage between a marker and a QTL for a trait. Charcosset and Essioux (1994) state that such a 'linkage disequilibrium', which they define as a statistical association between a marker and a QTL, should be similar in the parent groups of interest in order to predict heterosis succesfully. Using cross validation, Virk et al. (1996) show, for several agronomic and morphological traits in rice, that marker-trait associations in a stratified sample of genebank accessions are most likely based on linkage between markers and QTL. This would support the prediction of heterosis and progeny variance by genetic distances based on markers associated with traits in the population of parents. However, to confirm linkage, one needs to map QTL in offspring populations of these parents, which is rather laborious, or one could use published QTL-information from related parent combinations.

'Minor gene' variation is defined as genetic variation that cannot be explained by the segregation of 'major genes' that have been identified in QTL-analysis or segregation analysis. It is thought to be based on many segregating loci with small effects, which is a common assumption in biometrical genetics (Mather and Jinks, 1982). We expect that these loci are more or less uniformly spread over the genome. The amount of 'minor gene' variation for complex traits, like yield, can be quite substantial (see, for instance, Tinker et al., 1996). Exploitation of this variation can be performed after visual or marker-assisted selection for desired alleles of 'major gene' loci. The latter selection is often already performed during early inbreeding generations or during seed multiplication of DH lines. For the prediction of 'minor gene' variance, it is likely that parental relationship measures based on genetic differences that are evenly distributed over the genome, will perform well. Molecular markers do not always satisfy this demand. For instance, using AFLP markers in barley, Qi et al. (1998) found regions with high densities of markers, mainly positioned at the centromeric regions. Therefore, map information may be used to select regularly dispersed markers. Dillman et al. (1997) use a marker variance-

covariance matrix in the estimation of genetic distance to give more weight to markers from chromosome regions with low marker density.

In this study we investigate the prediction of barley progeny variance for some agronomic traits by parental genetic distance based on AFLP markers associated with these traits. Selected marker information is also used for the prediction of 'minor gene' variance. We use genetic map information to obtain a genetic distance that is weighted for marker density. The usefulness of marker selection for the prediction of progeny variance in practical plant breeding is discussed.

Material and methods

Plant materials

In this study we used twenty populations of 48 recombinant inbred lines (RILs) each. These were derived from crosses between 18 Northwest European two-row spring barley parents. Crosses and inbreeding procedure are described in chapter 4.

AFLP analysis

Genetic distances were based on 411 polymorphic AFLP markers originating from 14 primer combinations. Primer combinations and details about the AFLP analysis are given in chapter 4.

Field trials

The 18 parents plus two additional standards were tested at several locations and years in the years 1993 to 1996. The 960 RILs were tested in 1995 at two locations and in 1996 at five locations. We observed plant height (cm), flowering time (°C.days), thousand kernel weight (g) and yield (kg/ha). A more extensive description of the field trials is presented in chapter 4.

The details of the statistical analyses of the trials are also given in chapter 4. We used a linear mixed model with random RIL effects to estimate the genetic variance V per cross for every trait over environments (plant height, flowering time, thousand kernel weight, and grain yield). The generalised genetic variance GGV (Sokal, 1965), combining the variation for the differents traits, was calculated from the RIL data for each environment. We also calculated trait specific GGVs from single trait data for each environment. All GGVs were transformed by taking their n^{th} root, where n is the number of trait by environment combinations (see chapter 4).

Segregation analysis

In the RIL populations we observed 'major genes' segregation patterns for several traits. To establish significant segregation of a 'major gene' we fitted a bimodal mixture model ('major gene') and tested it against the unimodal model ('no major gene'). In this way we could postulate 43 'major genes'. A more detailed description of the analysis is presented in chapter 4.

We were able to estimate the 'minor gene' variance (mgV) by using postulated 'major genes', their mutual interactions and their interaction with the environment as explanatory variables in the analysis of variance. 'Minor gene' generalised genetic variances (mgGGVs) were calculated similarly as GGVs. They are based on RIL least squares means corrected for the effect of 'major genes' and mutual interactions of 'major genes'.

Map-based genetic distance

For 90 out of 1248 AFLP markers (non-redundant within primer combinations (see chapter 4) we were able to establish the map positions. For this we used the combined genetic map based on the genetic maps for the crosses Vada x L94 (Qi et al.,1998) and C-123 x L94 (de Bruin, unpublished). Out of these 90 mapped markers 65 were polymorphic among the 18 parents. The aim was to obtain a genetic distance that was based on a uniform representation of the genome. Therefore, a density dependent weight w_k (cM) for marker k was calculated as follows:

$$w_k = \frac{(mp_{k+1} - mp_{k-1})}{2\ n_k}$$

where mp_{k+1} is the map position in (cM) of the next marker on the chromosome, mp_{k-1} is the map position (cM) of the previous marker on the chromosome, and n_k is the number of markers mapped at the same chromosome position of marker k. When marker k was at an outer end of the map, mp_{k+1} or mp_{k-1} was replaced by mp_k.

After calculating the weights of the markers, the map-based genetic distance mgd_{ij} between parents i and j was calculated as follows: $mgd_{ij} = 1 - mgs_{ij} = 1 - \dfrac{2 \sum_{k=1}^{90} w_k\ p_{ijk}}{\sum_{k=1}^{90} w_k\ (p_{ik} + p_{jk})}$

where p_{ijk}=1 if marker k is present in both genotypes i and j,

p_{ijk}=0 if marker k is absent in at least one of the genotypes i and j,

and p_{ik}, p_{jk}=1 if marker k is present in genotype i, resp. j,

p_{ik}, p_{jk}=0 if marker k is absent in genotype i, resp. j

In case of a missing observation for a marker in genotype i and/or j, this marker was not included in the calculation of mgd_{ij}.

Marker-trait association

To trace markers associated with parent traits we performed an analysis of variance for each trait using data from all available environments. Because the two small-plot trials in 1994 did not reflect the large plot conditions in which the RIL yields were obtained, their yield data were not used to establish marker-trait associations. We tested 275 AFLP-markers for their association with each trait. These are all polymorphic markers, after removing redundant markers within and between the 14 primer combinations.

To test the degree of association between a marker and a trait we applied four linear models with increasing complexity. Model (1a) only involves an environment effect and an average allele substitution effect over environments for the investigated marker k. In model (1b) we added the possibility for an allele substitution effect to vary among environments, thus including marker by environment interaction in the model. Model (2a) is comparable to model (1a), but a randomly chosen marker k' is added before adding the investigated marker k to the model. The allele substitution effect for the investigated marker is nested within the allele substitution effect of the randomly chosen marker. One by one, all available markers are used as random marker k' in the model, except the marker under investigation. So, for every investigated marker k, 274 models are fitted, using each of the other markers as random marker k'. Then the P-values of the F-test for the investigated marker are averaged over all 274 models and used for marker selection. This approach probably includes part of the epistatic effects and prevents selection of markers that are loosely linked to QTL and that may show statistical colinearity with more tightly linked markers. It is expected to produce more reliable probability values to select markers associated with traits as compared to model (1a).To include part, a marker selection procedure is tried that uses a model with an extra marker additional to the marker under investigation. Model (2b) is comparable to model (2a), but we added the possibility for an allele substitution effect to vary among environments, as in model (1b). This resulted in the following models:

$$z_{il} = \mu + \lambda_l + \beta_k x_{ik} + E_{ilk} \qquad (1a)$$

$$z_{il} = \mu + \lambda_l + \beta_{k(l)} x_{ik(l)} + E_{ilk} \qquad (1b)$$

$$z_{il} = \mu + \lambda_l + \beta_{k'} x_{ik'} + \beta_{k(k')} x_{ik(k')} + E_{ilkk'} \qquad (2a)$$

$$z_{il} = \mu + \lambda_l + \beta_{k'(l)} x_{ik'(l)} + \beta_{k(k'(l))} x_{ik(k'(l))} + E_{ilkk'} \qquad (2b)$$

where z_{il} is the trait observation for parent i in environment l, μ is the general level parameter, λ_l is the environment parameter for environment l, x_{ik} is the indicator variable for the investigated marker k in parent i, and $x_{ik'}$ is the indicator variable for another randomly chosen marker k' ($k' \neq k$) in parent i. β_k, $\beta_{k(k')}$, $\beta_{k(l)}$, $\beta_{k(k'(l))}$ are allele substitution parameters for marker k. They are, depending on the model, allowed to differ for every marker genotype of marker k' and/or every environment l. $\beta_{k'}$, $\beta_{k'(l)}$ are allele substitution parameters for marker k', from which the second is allowed to differ for every environment l. $E_{ilkk'}$ is the residual.

Markers were selected on the basis of the P-value in the F-test. Different thresholds were applied (see below). Trait-dependent genetic distances *sgd* were calculated similarly as *gd* in chapter 4, on the basis of markers selected for each trait separately. Distance *sgd1* was based on marker selection using model (1a) or (1b) and *sgd2* was based on marker selection using model (2a) or (2b). *sgd*-distances based on all four traits (plant height, flowering time, thousand kernel weight, and grain yield) and aimed to predict GGV and mgGGV over all traits, were calculated by using all markers that were selected on the basis of at least one trait.

When fitting models (2a) and (2b) we calculated the fraction of residual variance explained by adding marker k to the model. This was done per trait by environment combination for each of the 275 × 274 *k-k'* constitutions. For each k the explained fractions were averaged over k'. From these average explained fractions per environment, the highest value was taken and used as a trait-specific weight for each marker. Weighted genetic distances *sgd3* were then calculated using markers selected by model (2a) or (2b).

For comparison of the models the number of markers selected by model (1a) was kept equal to the number of markers selected by model (2a). For this model an optimal threshold value was determined based on the highest correlation coefficient between *sgd* and progeny variance. Analogously to model (1a) marker selection, the numbers of markers selected by model (1b) P-values were set equal to the numbers selected by model (2b) P-values. The threshold for model (2b) was set at a value that gave approximately equal numbers of selected markers than the optimal threshold for model (2a).

To investigate the situation without segregating 'major genes', we discarded all markers which were strongly associated with the trait under investigation. The discarded markers are expected to be linked to the postulated major genes for the considered trait. The remaining AFLP-markers include markers with a significant, but less strong marker-trait association, and which are supposed to be linked to 'minor genes'. Such markers were selected and used in calculation of the genetic distances *sgd4*, *sgd5*, and *sgd6* (analogously to resp. *sgd1*, *sgd2*, and *sgd3*). We used models (1a) or (2a) for marker deletion depending on the model that was used for marker selection. An arbitrary threshold probability of 0.001 was used for P-values from model (2a). The number of markers discarded on the basis of P-values from model (1a) was kept equal to the number of markers discarded when applying model (2a) with threshold 0.001.

A bootstrap procedure (Efron & Tibshirani, 1993) taking 1,000 samples from the set of 20 parent combinations, was used to test whether the correlation coefficients between genetic distances and RIL variances were significantly larger than zero.

Results

Results of the field trials and the segregation analysis have been presented in chapter 2. Segregation analysis of five agronomic traits (plant height, flowering time, lodging, thousand kernel weight, grain yield) resulted in 43 postulated segregating 'major genes' distributed over 20 crosses.

Map-based genetic distance

The map positions of the 90 markers used in the calculation of *mgd* are not evenly distributed over the recombination map. Marker distribution patterns were similar to those of the original maps, presented by Qi et al. (1998), with high marker densities around the centromere positions. At 10 map positions more than one marker was present with an average of 2.3 markers per map position. At 6 of these positions markers showed different polymorphism patterns among parents. These markers were produced by different primer combinations. Markers generated by the same primer combination and mapped at the same position showed identical or complementary polymorphism patterns.

Correlations of *mgd* with variance estimates were poor and non-significant (Table 5.1). *mgd* did not show any increase in correlation with variance compared to the unweighted genetic distance (data not shown) based on the same 90 mapped markers.

Genetic distance based on marker-trait association

The predictive value of the *sgd*s for the different variance estimates was clearly dependent on the intensity of marker selection. Threshold values were varied to find an approximate optimum selection intensity, generally producing the highest correlation coefficients between genetic distances based on selected markers and progeny variances. The optimum selection intensities appeared to be different for the different models. However, for each trait the numbers of selected markers were usually in the same order of magnitude. Model (2a) had a consistent optimum at treshold value 0.005, with numbers of selected markers ranging from 18 (grain yield) to 36 (plant height).

Table 5.1. Correlation coefficients between AFLP-based parental divergence measures and progeny variances. 1−*gs*=unweighted genetic distance based on all markers; *mgd*=genetic distance based on 90 mapped markers weighted for marker density; *sgd1, sgd2*=unweighted genetic distance based on markers associated with the trait(s) given in the second column and selected using model (1a) and model (2a) resp., except for flowering time, for which markers are selected using model (1b) and (2b) resp.; *sgd3*, as *sgd2*, but weighted for average explained fraction of residual variance; *sgd4*/*sgd5*/*sgd6*, as *sgd1*/*sgd2*/*sgd3*, but only selected markers that are less strongly associated (P≥0.001) with the investigated trait in the second column; V/mgV=variance/'minor gene' variance analysed over environments; GGV/mgGGV=generalised genetic variance/generalised 'minor gene' variance, based on a specified set of *n* trait by environment combinations; V'/mgV'=variance/'minor gene' variance analysed per environment and correlation coefficients averaged over environments; h=plant height; f=flowering time; t=thousand kernel weight; y=yield; *=0.01<P<0.05; **=P<0.01

variance	traits	1−*gs*	*mgd*	*sgd1*	*sgd2*	*sgd3*	*sgd4*	*sgd5*	*sgd6*
V	h	0.31	0.20	0.44**	0.62**	0.59**	0.52**	0.76**	0.70**
mgV	h	0.43*	0.01	0.39	0.45**	0.49*	0.26	0.36	0.46*
V	f	0.03	0.20	0.19	0.19	0.26	-0.11	0.30	0.31
mgV	f	0.27	0.30	0.19	0.28	0.33	-0.26	0.50*	0.51*
V	t	0.09	0.17	0.48*	0.46**	0.48**	0.33*	0.44**	0.46**
mgV	t	0.02	-0.04	0.11	0.11	0.11	-0.04	0.13	0.13
V	y	0.02	0.21	0.09	0.38*	0.38*	0.23	0.40*	0.44*
mgV	y	-0.29	-0.13	0.33*	0.38*	0.34	0.44*	0.44*	0.42*
$GGV^{1/n}$	hfty	0.51**	0.24	0.47**	0.49*	0.54**	0.23	0.34	0.42*
$mgGGV^{1/n}$	hfty	0.47*	0.13	0.18	0.31	0.35	0.12	0.50*	0.57*
$GGV^{1/n}$	h	0.48*	0.14	0.62**	0.72**	0.68**	0.57**	0.82**	0.81**
$mgGGV^{1/n}$	h	0.39	0.11	0.25	0.34	0.34	0.14	0.28	0.33
$GGV^{1/n}$	f	0.27	0.18	0.22	0.21	0.28	-0.12	0.32	0.34
$mgGGV^{1/n}$	f	0.31	0.14	-0.07	-0.01	0.03	-0.29	0.34	0.34
$GGV^{1/n}$	t	0.48*	0.11	0.43*	0.61**	0.61**	0.46*	0.70**	0.71**
$mgGGV^{1/n}$	t	0.26	0.04	0.07	0.23	0.22	-0.06	0.36*	0.35*
$GGV^{1/n}$	y	0.13	0.02	0.01	0.23	0.30	0.15	0.24	0.33
$mgGGV^{1/n}$	y	0.34	-0.14	0.14	0.32	0.35*	0.37*	0.41*	0.45*

It appeared that models (1a) and (2a) with a general allele substitution effect over environments generally performed better than models (1b) and (2b). This was the case for plant height, thousand kernel weight, grain yield, and combined traits. However, for flowering time models (1b) and (2b), having an environment-dependent allele substitution effect, performed better than models (1a) and (2a). Therefore we decided to use models (1a) and (2a) for plant height, thousand kernel weight, grain yield, and combined traits and models (1b) and (2b) for flowering time for further comparisons (Table 5.1).

The single marker models (1a) and (1b) were compared to the models (2a) and (2b), that have the additional random marker (Table 5.1). The optimal threshold value of 0.005 was used for model (2a) P-values. In the investigation of flowering time, a threshold value of 0.05 was

used for model (2b) P-values. This threshold resulted in a number of selected markers that was close to the marker number that would have been selected using model (2a) with threshold 0.005. For comparison of the models the numbers of markers selected by models (1a) and (1b) were kept equal to the numbers of markers selected by model (2a) and (2b), as already mentioned in the material and methods section.

Correlation between genetic distance, based on selected markers, and progeny variance measures appeared to be generally higher than the correlation between genetic distance, based un unselected markers, and RIL variance estimates. The largest increase was found for correlations with trait variance V over environments. Marker selection on the basis of averaged P-values using model (2a) or (2b) resulted in higher correlations than marker selection based on model (1a) or (1b) P-values. The use of the average fraction of explained residual variance as a weight in the genetic distance calculation did not have a large effect on the correlation between genetic distance and progeny variance. Correlations with variance measures were more or less the same for *sgd2* and *sgd3*. In some cases there was an indication that weighting of markers had an increasing effect on the correlation coefficients.

In the situation without segregating 'major genes' the correlation coefficients between *sgd5* or *sgd6* and 'minor gene' variance for plant height and thousand kernel weight decreased compared to the situation with 'major genes'. For flowering time and yield an increased correlation was found. The correlation between *sgd5* or *sgd6* and V or GGV appeared to be equal or even higher than the correlation between *sgd2* or *sgd3* and 'minor gene' variances, except for the GGV for combined traits.

Considering the different traits, correlation coefficients between genetic distances and total RIL variances (V, GGV) were the highest for plant height, followed by thousand kernel weight. Variances for flowering time appeared hard to predict, as almost no significant correlations (α=0.05) were found. This was also the case for the GGV for yield. Marker selection hardly increased correlations between genetic distance estimates and generalised genetic variances for the combined traits.

Discussion

A genetic distance measure between parent lines, where map information was used to weight markers for the marker density at their position on the DNA, does not appear very successful in predicting progeny variance. This measure was an attempt to obtain a more balanced representation of the genome. As already mentioned by Charcosset et al. (1991), a highly predictive marker-based genetic distance should only involve markers that are linked to QTL for the specified trait. This means that only part of the genome should be represented in the genetic

distance measure. However, for the variance measure GGV, based on combined traits and presumably involving a high number of QTL, the correlation coefficient with *mgd* (r=0.24) is hardly higher than with 1−*gs*, based on the same 90 markers (r=0.21). As there is a discrepancy between the physical map (in Kbp), the recombination map (in cM) and the positions of expressed genes, it is unclear how a balanced genome representation should be obtained and whether it is representative for the GGV-QTL.

The genetic distance measures based on markers that are selected for their association with a certain trait, can be interpreted as agronomic distances (*agd*), presented in chapter 2, that are supported by molecular marker information. This explains the higher correlation coefficients of *sgd1*, *sgd2*, and *sgd3* with total variances V and GGV than with 'minor gene' variances mgV and mgGGV, showing a similar pattern as the correlation coefficients of *agd* in part I. In comparison with agronomic distances and genetic distances based on all available markers, *sgd* generally gives a higher correlation coefficent with trait variances. The increase was not found for flowering time variance (V) over environments. Closer inspection learned that a more stringent selection of markers could improve the predictive value of the *sgd* estimates for flowering time. For the other traits a threshold of 0.005 for model (2a) P-values seems to be a good selection criterion. It resulted in 18 (grain yield) to 36 selected markers (plant height) for single traits and 82 selected markers for the combined traits. It is, of course, unknown whether these marker selection intensities can be transferred to other situations.

The choice of the model to estimate type I error probabilities (P-values) probably depends on the type of marker by environment interaction. Considering the results in Table 5.1, models (2a) and (2b) are clearly preferred above models (1a) and (1b), but the question whether or not to include marker by environment interaction in the model needs more reasoning. If there is hardly any interaction (plant height, thousand kernel weight), model (2a) performs well, as expected. However, model (2a) also performs better than model (2b) for grain yield, for which marker by environment interaction is surely expected according to our segregation analysis results. This interaction partly consists of markers that have hardly any association with yield in one environment and strong association in the other environment. This is in contrast with marker by environment interaction for flowering time, where most markers are either always or never associated with the trait. The allele substitution effect is different, depending on sowing date, which causes the interaction. Thus the associated markers fit well to model (2b) and are important for the amount of variance over environments. For yield the markers with large interaction effects are not always relevant for the variance over environments. Model (2a) more or less ignores the interaction effects and is therefore more useful in selection of markers for yield. The use of weights for the selected markers is not advantageous, as the differences in correlation coefficients with RIL variances between weighted *sgd*s and unweighted *sgd*s are small.

The 'minor gene' variances were included in this study to resemble a situation which often occurs in practical barley breeding, as well as other crops: 'major genes' are already fixed at certain alleles in the closely related parent population or the segregating populations have been selected for desired alleles in the early generations of the inbreeding process. In the latter case it is advisable to use markers, closely linked to the 'major gene', as additional parameters in the marker-trait model. In this way the variation among parents, caused by the 'major genes', is removed before testing the association of other markers with the investigated trait. This is probably more reliable than the marker deletion procedure used for the calculation of *sgd4*, *sgd5*, and *sgd6*. However, we did not have sufficient information about the map positions of markers and 'major genes' to estimate a more reliable genetic distance.

The elimination of 'major genes' resulted in a strong decrease in the correlation coefficients of genetic distance measures and RIL variances for plant height and thousand kernel weight. These lower correlation coefficients give probably a more realistic view of the opportunities for prediction of progeny variance in practical breeding by genetic distances based on selected markers. We prefer to consider the correlation coefficients for yield, because of the consistent magnitude of the correlation coefficients, with and without 'major genes'. These correlations are mainly positive and significant. They seem to be higher than the correlations between genetic similarity and heterosis presented by Dudley et al. (1991). We doubt, however, whether the correspondence between *sgd* and progeny variance is high enough to be successfully applied for cross prediction in practical plant breeding.

Conclusion

The use of map-based marker densities as weights in the calculation of genetic distance does not improve its correspondence with progeny variance. This can be understood when one assumes that the effects on RIL variance of QTL segregation are not uniformly distributed over the linkage map.

The use of markers selected on the basis of their high degree of association with parental traits is therefore much more promising. The preferred marker selection is based on the average P-value obtained from model (2a) and (2b) where all remaining markers are included one by one, while testing the one under investigation. Depending on the trait, a model should be chosen that selects markers that are linked to trait effects which are relevant over environments.

The correlations between genetic distance *sgd*, based on selected markers, and progeny variance are mainly positive and range from poor to high. However, we found that high correlation coefficents are the result of a few segregating 'major genes'. These genes are often already fixed in practical breeding populations, either because of parent selection or because of selection

during early generations. Ignoring 'major gene' effects, we observe mainly poor-to-moderate correlations between *sgd* and RIL variance. This leads us to the conclusion that marker selection based on parental marker-trait associations, although promising, is not (yet) reliable enough to establish highly predictive genetic distance estimates for practical plant breeding.

Acknowledgements

The authors thank Cebeco Zaden B.V., Vanderhave Research and Zelder B.V. for making available some of their advanced breeding lines and extensive trial facilities. They thank CPRO-DLO and CIMMYT, Mexico for the use of trial facilities. They also thank Izak Bos, Johan Dourleijn, Johannes Kalma, Derk Gesink, XiaoQuan Qi, and Paul de Bruijn for their contributions to this study.

6

Prediction of barley progeny performance in the presence of genotype by environment interaction [5]

Johan W. Schut and C. Johan Dourleijn

Abstract

Twenty crosses of European two-row spring barley and their parents were tested in six environments in the Netherlands to investigate the inheritance of genotype by environment interaction. First, the inheritance of three stability measures is considered: Finlay and Wilkinson's (1963) regression coefficient b_i, Shukla's (1972) stability variance σ_i^2 and Eberhart and Russell's (1966) mean squared deviation d_i^2. The average b_i value of the offspring recominant inbred lines (RILs) is strongly correlated with the midparent value, indicating its heritable nature. The correlation between RIL mean and midparent value is absent for σ_i^2, due to a difference in its composition for parents and RILs. d_i^2 appeared to be heritable. However, its repeatability is poor. Therefore, it is concluded that only prediction of b_i is useful in practical plant breeding.

Secondly, a biplot from the AMMI-analysis of the parents is investigated. RIL means are added in the biplot. The first two axes represent b_i and the difference between clay and sandy soils. There appears to be reasonable correlation between the RIL mean positions and the midparent positions in the biplot. However, a midparent prediction of offspring genotype by environment interaction, based on the AMMI-biplot, is probably not reliable enough for practical purposes.

Finally, Habgood's (1977) parental similarity measure is calculated as the correlation between the parental residual vectors from a two-way ANOVA of the parent by environment table using a model with additive genotype and environment effects. It shows a reasonable negative correlation (r_s=−0.63) with offspring variance for yield over environments. It is concluded that the use of this similarity measure to predict progeny variance in practical plant breeding appears promising, but further investigation is necessary.

Keywords: Additive Main effects and Multiplicative Interaction, genetic similarity, *Hordeum vulgare*, progeny variance, stability

[5]submitted for publication

Introduction

Genotype by environment interaction is a very important phenomenon in practical plant breeding. A breeder can select cultivars specifically adapted to a certain location, but more often he prefers cultivars that show a stable yield performance over several years and locations. As a breeding goal this stability is often just as desirable as a high yield level. In general a genotype is regarded as stable when its performance across environments does not deviate from the average performance of a group of standard genotypes. Several measures have been presented to quantify this feature. Extensive reviews are presented by Lin et al. (1986) and Becker and Léon (1988).

Lin et al. (1986) considered three types of stability parameters. Type 1 stability is accomplished when a genotype shows a small variance over environments. A genotype shows type 2 stability, when its performance in a certain environment can be predicted by an additive model consisting of a genotype term and an environment term. The latter term is equal to the average yield over all genotypes in a certain environment (i.e., the environmental index). An example of a type 2 stability parameter is the regression coefficient from the regression of the yield of genotype i on the environmental index (Finlay and Wilkinson, 1963). Eberhart and Russell (1966) consider a value of one, i.e., average stability, as a desirable value of b_i. However, originally Finlay and Wilkinson proposed a value of zero as the desirable value. In that case Lin and Binns (1991) propose to classify b_i as a type 1 stability parameter. Another example of a type 2 stability statistic was proposed by Shukla (1972). Based on the residuals from the additive model this stability variance σ_i^2 of genotype i is defined as the variance of a genotype across environments. Type 3 stability is based on the deviation from the Finlay-Wilkinson regression. Eberhart and Russell (1966) proposed to use the mean squared deviation from the regression (d_i^2). A genotype is considered stable when this parameter is zero. Becker and Léon (1988) showed the relationship between b_i, σ_i^2, and d_i^2, where σ_i^2 is the sum of a linear term based on b_i and a non-linear term d_i^2. Later, Lin and Binns (1988) defined type 4 stability, which is merely a modification of type 1 stability, ignoring the variance among locations. A genotype is considered type 4 stable when its performance does not vary over years.

The repeatability over different sets of environments and the genetic control of several stability parameters (e.g. b_i, σ_i^2, d_i^2) is reviewed by Sneller et al. (1997). Based on reviewed literature they conclude that the repeatability of these parameters was generally low for σ_i^2, d_i^2 and moderate for b_i. The regression parameter b_i is reported to be under genetic control, but σ_i^2 and d_i^2 do not seem to be heritable. Using soybean grain yield data, measured in several environments over two years, Sneller et al. (1997) could confirm the earlier conclusions about repeatability of stability parameters. Lin and Binns (1991) make general statements about the genetic control of the different types of stability parameters. On the basis of bromegrass forage

yield data, measured at four locations in three years, they conclude that only type 1 and type 4 statistics are heritable. This seems to be in contradiction with the conclusion of Sneller et al. (1997) that the regression parameter b_i is under genetic control, although it may be dependent on crop and/or trait.

Genotype by environment interaction is usually investigated by inspection of the deviations from the two-way ANOVA model with additive genotype and environment effects. Gollob (1968) proposed a factor analysis for parsimonious modelling of this non-additivity by means of a few multiplicative terms. This bilinear model is also known as Additive Main effect and Multiplicative Interaction (AMMI) model (Zobel et al., 1988). It is expected that the interaction pattern of parent lines is an indication for the interaction pattern of the offspring. Van Eeuwijk (1995) mentions that the cosines of the angles between genotypic vectors approximate the correlations between genotypes with respect to their interactions with environments. Habgood (1977, 1983) proposed to use these correlations to estimate genetic diversity among parents to obtain an indication for the variation in yield among the offspring. This is based on the idea that genetic similarity of genotypes with respect to yield genes will result in a similar response to environmental changes, and, therefore, in high correlations between the genotype vectors. Conversely, genotypes with non-similar responses will have few yield genes in common. Therefore, the similarity of reaction patterns of two genotypes might be related to the yield variance among the segregating offspring: a higher similarity will result in less variation among the offspring.

To investigate the inheritance of genotype by environment interaction, and stability in particular, 20 crosses of European two-row spring barley and their parents were tested in six environments. We consider the genetic properties of three stability measures: Finlay and Wilkinson's (1963) regression coefficient, Shukla's (1972) stability variance and Eberhart and Russell's (1966) mean squared deviation. We will discuss the opportunities of AMMI analysis of parent genotypes to predict genotype by environment interaction for their offspring. We will also consider the prediction of yield variance among the offspring of a cross based on the correlation between the parental environment-response vectors.

Material and methods

Plant materials

For this study we used 20 populations of recombinant inbred lines (RILs) of two-row spring barley (*Hordeum vulgare* L.). These populations were derived via a partial diallel crossing design using 18 parent genotypes (n=18; s=2; Kempthorne and Curnow, 1961) and resulting in 18 crossing combinations (Table 6.1). Reciprocal crosses were made for two parent

combinations, increasing the total number of RIL populations to 20. The parent genotypes represent parents employed in commercial barley breeding programs in Northwest Europe over the last 20 years. Their pedigree and geographic origin are presented in Table 2.1. Single seed descent was performed on 48 F2 plants for each cross until the F5 generation. The F5 plants produced, after one intermediate generation of multiplication, a total of 960 F5-derived F7-lines. These recombinant inbred lines could then be tested and compared with their parents.

Table 6.1. Crosses and their parents. R=reciprocal combination

cross	mother	father	cross	mother	father
1	Riff (1)	Drossel (10)	11	Karat (14)	Yriba (13)
2	Baronesse (3)	Forester (2)	12(R2)	Gunhild (15)	GEI-119 (5)
3	Baronesse (3)	Bonaire (16)	13	Gunhild (15)	CEB-9186 (17)
4	Apex (8)	Riff (1)	14	Bonaire (16)	Porthos (9)
5	Porthos (9)	Yriba (13)	15	CEB-9186 (17)	ZE-87-3414 (4)
6	Midas (12)	Forester (2)	16	ZE-87-3414 (4)	CEB-9079 (18)
7	GEI-119 (5)	Midas (12)	17(R2)	GEI-119 (5)	Gunhild (15)
8(R1)	Prisma (6)	Apex (8)	18	Triangel (7)	Georgie (11)
9	Prisma (6)	Karat (14)	19(R1)	Apex (8)	Prisma (6)
10	Triangel (7)	Drossel (10)	20	Georgie (11)	CEB-9079 (18)

Field trials

The 960 recombinant inbred lines (RILs) were tested at several locations in the Netherlands in 1995 and 1996 (Table 6.2). In 1996, three trials (Lelystad, Rilland, and Ottersum) each contained only one third of the lines (i.e., 16 lines) of each cross. They are called 'partial' locations, while the other trials are indicated as 'complete' locations. The 18 parent lines and cultivars Magda and Vada were added as standards. Each trial included two replicates. Each standard occurred six times per replicate at the 'complete' locations and two times per replicate at the 'partial' locations. All genotypes were randomised according to a partially balanced incomplete block design with 8 plots per block. The constraint that two genotypes do not occur more than once together in the same block, extended over all RIL trials. Sowing dates and plot sizes are presented in Table 6.2. Grain yield was recorded in grams per plot and recalculated to kg/ha.

Table 6.2. Field trial description. p=present in trial; †=only part of RILs present in trial

trial	location	soil	year	plot size (m^2) (width × length (m))	sowing date	parents	RILs
S95	Swifterbant	clay	1995	9.0 (1.5 × 6.0)	15 May	p	p
W95	Wageningen	clay	1995	9.0 (1.5 × 6.0)	11 May	p	p
S96	Swifterbant	clay	1996	9.0 (1.5 × 6.0)	1 April	p	p
W96	Wageningen	clay	1996	9.0 (1.5 × 6.0)	19 March	p	p
L96	Lelystad	clay	1996	5.32 (1.4 × 3.8)	18 April	p	p†
R96	Rilland	clay	1996	3.6 (1.5 × 2.4)	19 March	p	p†
O96	Ottersum	sand	1996	4.65 (1.5 × 3.1)	18 March	p	p†

Statistical analysis

Field trials

RIL-trial data were analysed using average information REML (Gilmour et al., 1995). Analysis of the plot data was performed per year by location combination (environment), because an overall analysis appeared not feasible due to computational limitations. The linear mixed model included fixed effects for standards, crosses and strips of adjacent incomplete blocks. The block effects were assumed random, as well as the line within cross effects. In the analysis of residuals and whenever a hypothesis considered the specific lines that were present in the trial, the line effects were assumed to be fixed. Residual analysis was performed to trace outliers among the data. These observations were excluded from the final analyses.

To overcome computational limitations we performed a combined analysis over years and locations (environments) by using the least squares means for the lines as input data for an analysis of variance. The linear mixed model included fixed effects for standards, crosses, standard by environment interaction and cross by environment interaction. A random line effect over environments was included to obtain an estimated line variance V for yield for each cross.

Stability parameters

An index value for each environment was calculated as the mean yield of all parent genotypes. Regression of the RIL and parent data on this environmental index produced Finlay and Wilkinson's (1963) regression coefficient b_i and the mean squared deviation from the regression d_i^2 (Eberhart and Russell, 1966) for each genotype i. Residuals were derived from the analysis of variance of least squares means data using a model with additive genotype and environment effects. Based on these residuals we calculated the stability variance σ_i^2 for an individual genotype i, following Shukla (1972). The stability variances of the RILs were calculated separately from those of the parents.

Using the individual RIL statistics, means and variances for the stability parameters were estimated per cross. The influence of high inaccurate values of d_i^2 and σ_i^2 was decreased by a log-transformation before calculating mean and variance.The inverse error variance of b_i and the degrees of freedom for d_i^2 and σ_i^2 were used as weights in the calculation of mean and variance.

AMMI analysis

The least squares means from the analysis of variance of the individual locations were used to construct the parent genotype by environment table for the AMMI analysis. Additive genotype and environment parameters were fitted by analysis of variance. The nonadditive part was described by principal components analysis. Sensitivity scores were recorded for every parent genotype. Results of the parental AMMI-analysis were visualised by a biplot.

Next the RIL data were inserted in the biplot as follows. A mixed model including fixed cross and environment effects, fixed cross by environment interaction, and a random line within cross effect was used to produce best linear unbiased predictions for the missing RIL by environment combinations (Van Eeuwijk, 1995). Then the environment parameters estimated from the parent data were used to eliminate the environment effect from the RIL data. Subsequently additive genotype effects were separated from the nonadditive part of the RIL data by analysis of variance. The nonadditive part was transformed into sensitivity vectors by a linear combination given by the eigenvectors from the parental AMMI-analysis. Parent and offspring could then be compared in the biplot.

Finally, correlations $r_{G\times E}$ between the parental environment-response vectors were calculated following Habgood (1977), and compared with RIL variance V for yield and stability parameters. The parental environment-response vectors consist of the residuals of an analysis of variance (ANOVA) of the parent by environment table using a model with additive genotype and

Table 6.3. Field trials characterised by average yield, observed over 20 crossing populations and 20 standards and root mean square errors (root mse) obtained by variance analysis.

trial	mean yield (kg/ha)	√mse (kg/ha)
S95	5452	229
W95	6048	220
S96	-	-
W96	9127	293
L96	7404	301
R96	9813	453
O96	9067	261

environment effects. The correlation $r_{G\times E}$ is equal to the cosine of the angle γ_{ij} between the two parental sensitivity vectors from an AMMI analysis using all dimensions. In this analysis the genotype scores for the different dimensions are multiplied with the singular values for those dimensions (Van Eeuwijk, 1995). It is possible to discard the highest dimensions to obtain a more parsimonious model and maybe, to lose some random non-genetic variation. Then the cosine of γ_{ij} is still a good approximation of the correlation $r_{G\times E}$. We investigated the effect of this dimension reduction on the correlation between $\cos(\gamma_{ij})$ and progeny variance V.

Results

Field trials

Average yield values and root mean square errors per trial are presented in Table 6.3. Mean yields of the parents are presented in Table 6.4. For each cross mean yield over environments and the square root of the estimated between line variance are presented in Table 6.5. Due to severe hail storm damage, we did not obtain yield data from Swifterbant in 1996.

Table 6.4. Average yield (kg/ha) over environments and stability parameters of parents. b_i=regression coefficient (Finlay and Wilkinson, 1963); $\ln(d_i^2)$=natural logarithm of mean squared deviation from regression (Eberhart and Russell, 1966); $\ln(\sigma_i^2)$=natural logarithm of stability variance (Shukla, 1972)

parent	name	yield	b_i	$\ln(d_i^2)$	$\ln(\sigma_i^2)$
1	Riff	8211	0.97	11.80	11.79
2	Forester	7810	1.09	11.42	11.74
3	Baronesse	8428	1.13	12.58	12.74
4	Ze-87-3414	7526	1.08	12.56	12.58
5	GEI-119	8136	0.75	12.23	12.97
6	Prisma	8072	0.97	11.04	11.12
7	Triangel	8588	1.30	11.32	12.98
8	Apex	7846	0.98	11.59	11.59
9	Porthos	7287	0.89	11.42	11.83
10	Drossel	7526	0.98	10.69	10.84
11	Georgie	7769	0.97	9.40	10.02
12	Midas	7441	0.94	11.51	11.64
13	Yriba	7691	0.87	10.45	11.60
14	Karat	7784	0.96	11.26	11.37
15	Gunhild	7868	0.92	10.20	10.92
16	Bonaire	8109	1.06	11.93	11.97
17	CEB-9186	8133	1.05	11.21	11.37
18	CEB-9079	8276	1.07	11.85	11.96

Table 6.5. Cross population mean and square root of between RIL variance for yield over environments (kg/ha) and stability parameters (see Table 6.3).

			mean			√variance		
cross	yield	√var(yield)	b_i	$\ln(d_i^2)$	$\ln(\sigma_i^2)$	b_i	$\ln(d_i^2)$	$\ln(\sigma_i^2)$
1	7865	397	0.96	11.36	11.76	0.15	1.26	1.02
2	7857	292	0.97	10.96	11.58	0.15	1.63	1.19
3	8036	252	0.97	11.14	11.37	0.11	1.28	1.17
4	7842	254	0.95	11.19	11.64	0.15	1.11	0.93
5	7552	197	0.84	10.37	11.28	0.12	1.51	1.04
6	7350	461	0.93	11.24	11.58	0.15	1.21	1.13
7	7884	232	0.84	11.20	12.02	0.14	1.66	0.70
8	7659	357	0.94	10.94	11.65	0.15	1.43	1.05
9	8010	147	0.99	11.24	11.53	0.11	1.22	1.03
10	7953	199	0.98	11.04	11.12	0.10	1.12	1.06
11	7485	393	0.91	10.97	11.40	0.11	1.43	0.97
12	7798	153	0.87	11.15	11.50	0.12	1.39	1.09
13	7731	234	0.95	11.01	11.30	0.11	1.55	0.99
14	7671	356	0.93	10.98	11.34	0.12	1.54	1.22
15	7487	289	0.98	11.35	11.69	0.14	1.21	1.13
16	7578	336	0.97	11.07	11.63	0.12	1.20	0.85
17	7746	217	0.87	11.27	11.60	0.13	1.52	1.20
18	7922	241	0.96	10.91	11.24	0.09	1.17	0.70
19	7622	344	0.95	11.08	11.78	0.15	1.17	0.86
20	7805	492	0.91	10.65	11.23	0.15	1.62	1.20

Stability parameters

Stability parameters for the parents are presented in Table 6.4. Approximate standard errors were estimated using the jackknife method (Tukey, 1958). For b_i they ranged from 0.013 to 0.073 with an average of 0.037. For $\ln(d_i^2)$ standard errors ranged from 0.22 to 1.34 with an average of 0.52 and for $\ln(\sigma_i^2)$ the range was between 0.10 and 0.76 with an average of 0.32. The ratio between the standard deviation among the parents and the standard error was 3.23 for b_i, 1.55 for $\ln(d_i^2)$, and 2.42 for $\ln(\sigma_i^2)$. Mean and root variance of the different stability parameters for each RIL population are presented in Table 6.5.

The rank correlations between the midparent values and RIL means for the investigated stability statistics are presented in Table 6.6. The midparent value of b_i appears to be a good predictor of the average b_i value of the offspring. For $\ln(d_i^2)$ there is also a significant correlation between midparent value and RIL mean. For $\ln(\sigma_i^2)$ a relation between parent and offspring is completely absent. However, the midparent value of $\ln(d_i^2)$ appears to be moderately correlated with the RIL mean of $\ln(\sigma_i^2)$.

Table 6.6. Spearman rank correlation coefficients between midparent values and RIL means of stability parameters (see Table 6.3). *=0.01<P<0.05 **=0.001<P<0.01 ***=P<0.001

		midparent value		
		b_i	$\ln(d_i^2)$	$\ln(\sigma_i^2)$
RIL mean	b_i	0.69***	0.26	0.05
	$\ln(d_i^2)$	-0.28	0.39*	0.20
	$\ln(\sigma_i^2)$	-0.31	0.58**	0.17

AMMI analysis

The first two factors from the AMMI analysis are used to construct the biplot presenting parents, RIL population means and environments (Figure 6.1). These first two dimensions explain 69% of the nonadditive variation among the parents. The first dimension seems to present the difference in genotype response between Rilland trial in 1996 (R96) and the 1995 environments (S95, W95) plus the Ottersum trial in 1996 (O96). The second dimension mainly presents the difference between the response for the Ottersum trial in 1996 and the other environments.

The average positions of the RILs for each cross are generally closer to the origin of the biplot than their parents. Their positions tend to be more towards lower yielding environments (S95, W95, and L(elystad)96). A comparison of the average position of the RILs with the positions of their parents does not show a consistent relationship between them. However, the rank correlation coefficient between midparent coordinates and cross coordinates is 0.82 for the first dimension and 0.70 for the second dimension. In most cases the average offspring position is situated somewhere between the parents, e.g. cross 1 (1×10). But in some cases the average offspring position is not even near the parents, e.g. cross 2 (3×2).

The correlation $r_{G\times E}$ between the parental environment-response vectors was calculated for each of the 20 crosses. The correlation coefficients of $r_{G\times E}$ with between RIL variance for the investigated stability statistics were small and insignificant.The relationship between $r_{G\times E}$ and variance V for yield over environments is presented in Figure 6.2. The rank correlation coefficient between $r_{G\times E}$ and V is −0.63 (P<0.001).

Discussion and conclusion

A large variation in average yield was found among the investigated environments. The 1995 trials were sown late. Ripening was promoted by hot weather during the end of the season, especially in Wageningen. In 1996 the trials in Wageningen, Rilland and Ottersum were sown early; the Lelystad trial was sown on an intermediate date. The grain filling stage was largely extended as a result of cool weather, especially in Rilland and Lelystad. Severe lodging occurred

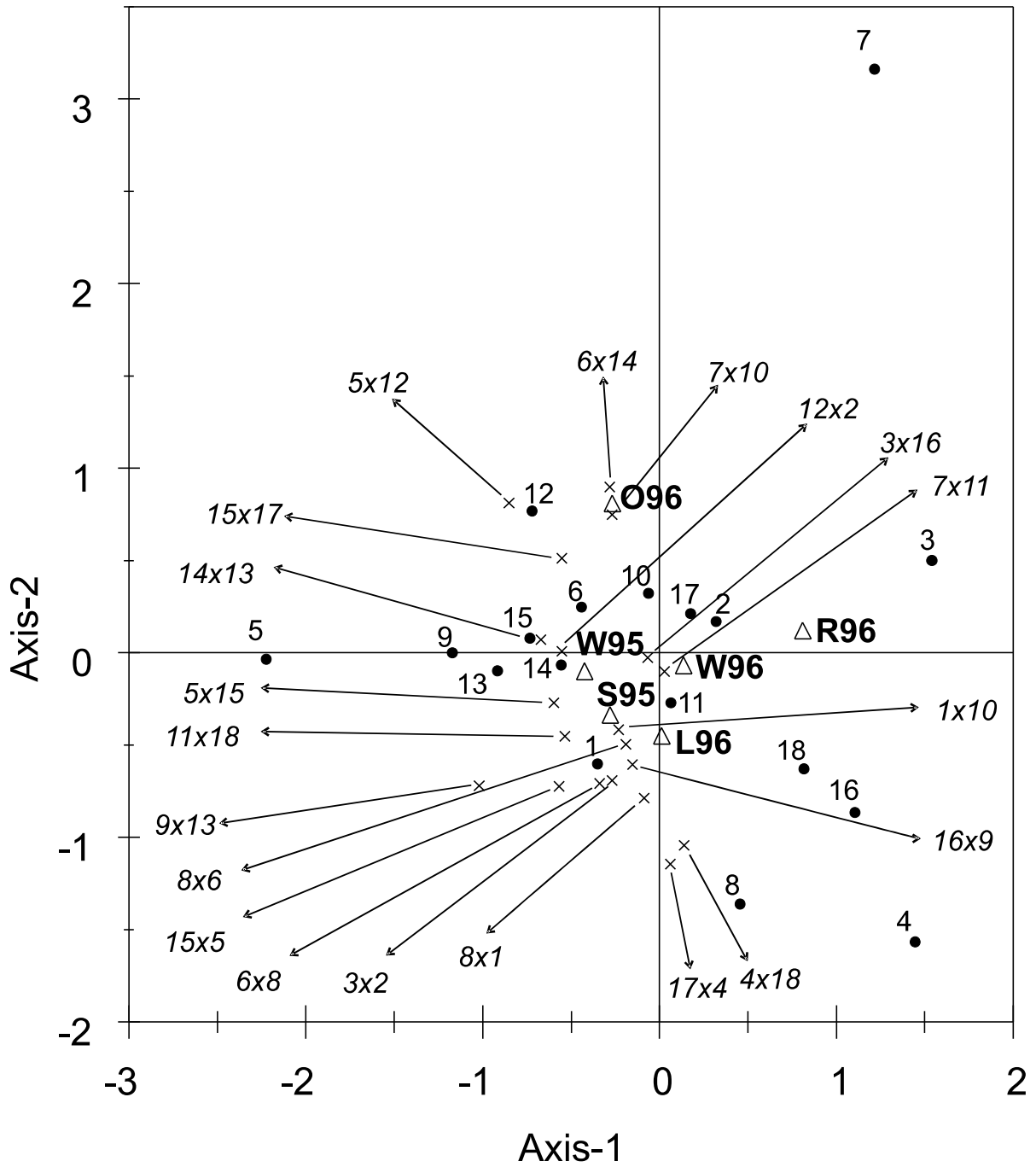

Figure 6.1. AMMI biplot of genotype by environment interaction for parents and recombinant inbred line (RIL) populations. ● = parents; × = crosses, presented as parent combinations; Δ = environments; L96=Lelystad 1996; O96=Ottersum 1996; R96=Rilland 1996; S95=Swifterbant 1995; W95/W96=Wageningen 1995/1996

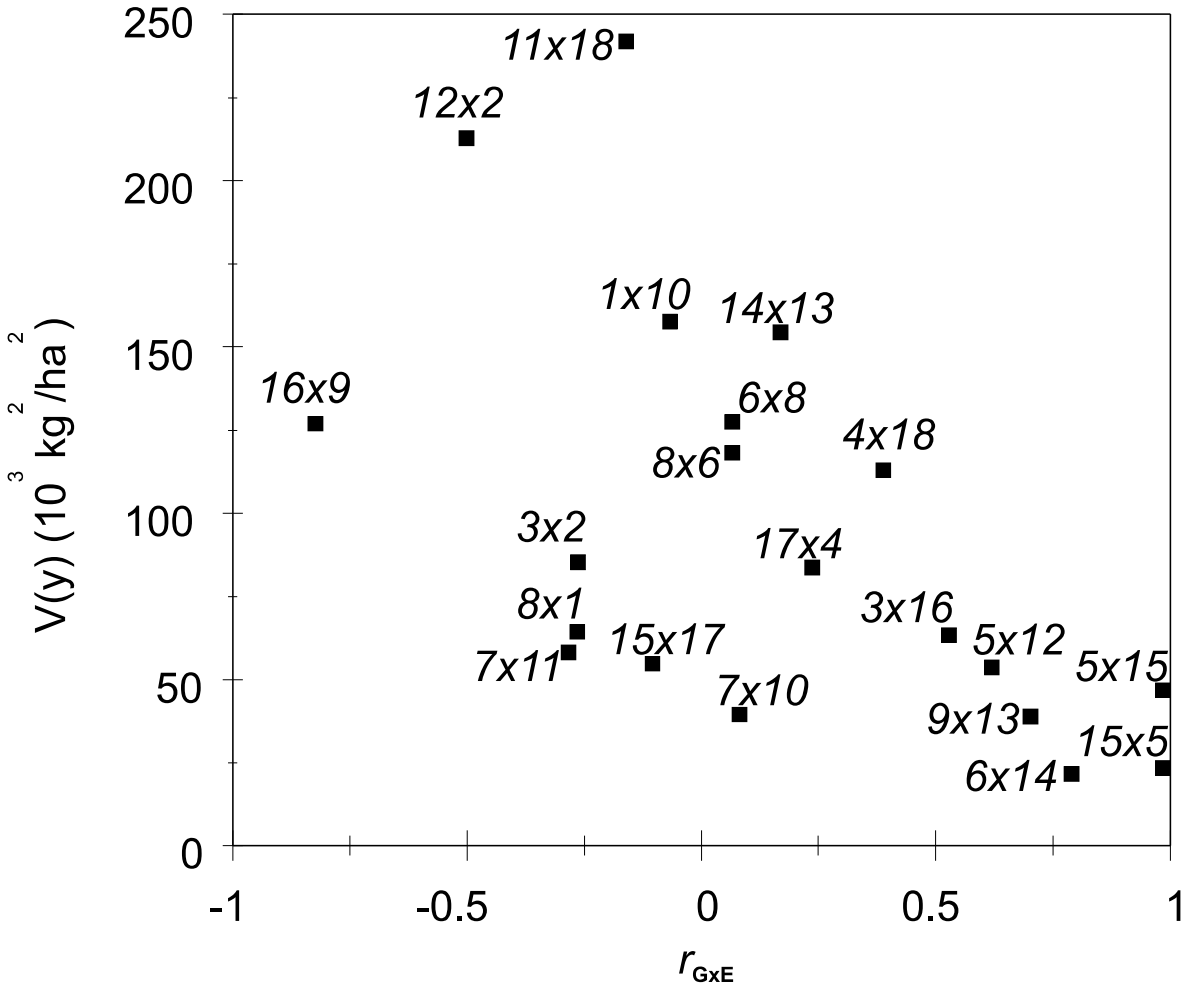

Figure 6.2. The relationship between the correlation coefficient $r_{G\times E}$ between environment-response vectors of two parents and the yield variance V(y) over environments among recombinant inbred lines descending from a cross between these parents. The environment-response vectors consist of the residuals of an analysis of variance of parent by environment data for yield using a model with additive genotype and environment effects.

in Rilland, Wageningen and Lelystad in 1996, and in Swifterbant in 1995. In the Wageningen trial in 1995 some lodging occurred towards the end of the season, and in Ottersum in 1996 there was hardly any lodging. The sowing date effect, which seems confounded with the year effect, could be an explanation for the differences in average yield between environments. The main indication for this is the yield difference between Lelystad and the other locations in 1996.

Investigation of the stability statistics of the parents shows a significant correlation between regression coefficient b_i and average yield of parent i (r_s=0.42), which was already mentioned by Finlay and Wilkinson (1963). A highly positive correlation (r=0.87) is observed

between the standard deviation among the parents and the average yield per environment, indicating that the differences between parents increase as the average yield of an environment increases. The positive rank correlations between parent performances in different environments (average r_s=0.46) show that the rank order of the parents does not change dramatically over environments. The combination of these two facts is in agreement with the correlation between b_i and average parent yield. Most high yielding genotypes perform relatively well in high yield environments and therefore have a high b_i value, while poor yielding genotypes more often have a low b_i value. The positive correlation between b_i and yield over environments is an indication that the effects of some of the yield QTL increase when moving from low to high yield environments.

The high correlation between the midparent value and the RIL population mean for b_i corroborates the conclusions of Becker and Léon (1988) that this stability measure is highly heritable. This would be in contradiction with the conclusion of Lin and Binns (1991), stating that type 2 stability parameters, to which they initially assigned b_i (Lin et al., 1986), are not inherited. However, in their discussion they reason that b_i can be interpreted as a heritable type 1 stability parameter as well.

For the second type 2 stability parameter, $\ln(\sigma_i^2)$, no significant correlation is found between midparent value and RIL population mean. This is an indication that this parameter is not heritable, which confirms the conclusion of Lin and Binns (1991). However, the midparent value of the natural logarithm of the mean squared deviation from the regression $\ln(d_i^2)$ (Eberhart and Russell, 1966) appears to be significantly correlated with the RIL mean for $\ln(\sigma_i^2)$ as well as with the RIL mean for $\ln(d_i^2)$. This can be interpreted as an indication that these type 2 and type 3 stability statistics do have a heritable component. As mentioned in the introduction, σ_i^2 is the sum of a linear term based on b_i and a non-linear term based on d_i^2 (Becker and Léon, 1988). The non-linear term is the more important component of σ_i^2. This is confirmed by the rank correlations between the parameters ($r_s(\ln(\sigma_i^2),\ln(d_i^2))$=0.81 and $r_s(\ln(\sigma_i^2),b_i)$=0.34 for parents; $r_s(\ln(\sigma_i^2),\ln(d_i^2))$=0.62 and $r_s(\ln(\sigma_i^2),b_i)$=−0.04 for RIL means). The difference between parents and RILs in the relative contributions of the two terms to σ_i^2 may be the main cause for the lack of relationship between its midparent value and the RIL population mean. Apparently the division of nonadditivity into a linear component b_i and a non-linear component d_i^2 clarifies its heritable basis. The relatively large inaccuracy of the d_i^2 estimates, especially for the RILs, may have had a decreasing effect on the correlation between midparent value and RIL mean. The contradiction of our results with the conclusions of Lin and Binns (1991) based on forage yield in brome grass, stating that type 2 and 3 stability parameters are not heritable, can probably be explained on the basis of differences in crop characteristics. The relationship between parent and offspring can be shown much clearer by using pure lines of a self fertilising crop, like barley, than by using a cross fertilising crop like brome grass. This is due to the absence of dominance

effects in the first situation. However, the jackknife standard errors indicate that the repeatability of the stability statistics σ_i^2 and d_i^2 is not very high, confirming the conclusion of Sneller et al. (1997). It is therefore questionable whether prediction of these stability statistics for offspring in one set of environments on the basis of parents in another set of environments will be useful in practical plant breeding.

Investigation of the biplot showed that the first principal component *pc1* almost coincides with the linear component of the genotype by environment interaction ($r_s(pc1,b_i)$= 0.96). The average b_i value of the RILs appears to be smaller than the average b_i value of the parents. In the biplot we see this in the average position of the RILs which is left of the origin and close to the environments with poorer yields. The positive correlation between b_i and average yield, observed for parents, can be confirmed by comparing parents and RILs. The average yield as well as the average b_i value of the RILs appear to be smaller than the average yield and average b_i value of the parents. The second dimension of the biplot is indicating the difference between the environment with a sandy soil and the environments with clay soils. This difference is somewhat confounded with the effects of lodging, which was much more severe at the clay environments. The positive correlation between the average coordinates of the parents and the coordinates of the RIL population mean indicates that there may be possibilities to use environmental sensitivities of the parents to predict the genotype by environment interaction patterns of the offspring. This is especially the case for the first dimension, the linear component, for which we already showed its heritable nature. However, comparison of parent and average offspring positions in the biplot shows that perspectives for prediction are limited.

We tried to predict the variance of the stability parameters in a RIL population from the parental sensitivities calculated in the AMMI analysis. Euclidean distance measures (data not shown) as well as $r_{G\times E}$ appear to be poor predictors. We assume that the inaccuracy of the variance estimate is too high to be able to find good predictors for stability parameter variance.

The between RIL variance V for yield over environments is negatively correlated with the correlation $r_{G\times E}$ between the parental residual vectors from a two-way ANOVA of the parent by environment table using a model with additive genotype and environment effects. This confirms results of Habgood (1983) showing that F2-populations from 'similar' parents (high $r_{G\times E}$) show less variation than F2-populations from 'dissimilar' parents. In contrast to the study of Habgood (1977; 1983) in which similarities were based on 40 environments, we only used 6 environments and already obtained a reasonable correlation with progeny variance. We suppose that the residual vectors represent QTL by environment interaction for yield. Different directions of the residual vectors represent different QTL with different QTL by environment patterns. The variance of the effect of a single QTL across environments is assumed to be positively correlated with the average size of the effect of that QTL over environments. Deviations from this relationship may cause differences in V-value between RIL populations with the same $r_{G\times E}$-

values. Also inaccuracies in the estimation of V and in the estimation of the least squares means that make up the parent by environment table cause a weakening of the correlation.

The correlation $r_{G\times E}$ is equal to the cosine of the angle γ_{ij} between the two parental sensitivity vectors from an AMMI analysis (Van Eeuwijk, 1995). The effect of dimension reduction in the AMMI-analysis on the correlation between $\cos(\gamma_{ij})$ and progeny variance V is investigated. We observed the strongest correlation between V and $\cos(\gamma_{ij})$ when we retained all five dimensions after AMMI analysis (data not shown), although the fifth principal component only explained 5% of the nonadditive variation. Inspection of this component, after standardisation, showed a rather extreme value of −2.88 for parent 17, indicating that this dimension is probably genetically meaningful and represents a genotype with a deviating genotype by environment interaction compared to the other parents. However, it is difficult to give a general indication of the number of dimensions that needs to be retained to determine γ_{ij} for variance prediction.

A comparison of $r_{G\times E}$ with other variance predictors, presented in chapter 2, showed no correlation with AFLP-based genetic distance and with a distance measure based on morphological characters. The correlation coefficient between $r_{G\times E}$ and the well-known coefficient of coancestry, based on pedigree data, was 0.42 (n=10). In chapter 2 it was suggested to combine several parental divergence measures to improve the prediction of V. However, none of the three distance estimates could explain the residuals from a regression analysis of V on $r_{G\times E}$. On the basis of the correlation between $r_{G\times E}$ and V, we conclude that the use of $r_{G\times E}$ for the prediction of progeny variance appears promising and encourages further investigation.

Acknowledgements

The authors thank Cebeco Zaden B.V., Vanderhave Research and Zelder B.V. for making available some of their advanced breeding lines and extensive trial facilities. They thank CIMMYT, Mexico for the use of trial facilities. They also thank Piet Stam, Izak Bos and Fred van Eeuwijk for their contributions to this study.

7

General discussion

In plant breeding programmes that create novel genetic variation by crossing, the choice of parent combinations is very important. A breeder would like to select those parent combinations that will produce the best performing offspring, as defined in chapter 1. To support the decision on the choice of parent combinations one could make use of knowledge about the genes underlying traits as well as knowledge about empirically established relationships between parents and offspring. In the present study cross prediction methods using the latter type of knowledge, and designated as B and D in chapter 1, are investigated in crosses between European two-row spring barley lines. We mainly consider the opportunities for prediction of offspring mean and variance.

The environments that are used to assess parent and cultivar performance represent barley growing conditions in the central and southwest part of the Netherlands. It can be seen from average yields and the biplot in chapter 6 that the trials are quite different and do not produce much redundant information. The locations Lelystad (1994 and 1996) and Swifterbant (1995) are in the same region, the central clay area of Flevoland, and represent similar growing conditions. In commercial barley breeding the test locations usually cover a larger region of Northwest Europe.

The data are analysed per environment using average information REML (Gilmour et al., 1995). Least squares means for the lines are used as input for an analysis over environments. In this analysis, crosses, environments and cross by environment interaction are assumed fixed, in order to retain maximum information for further correlation analyses. In these analyses we assume that parents and crosses are a more or less random sample from a larger candidate parent and cross population. The genetic variances are estimated per cross by assuming random line effects. For the situation of line prediction fixed line effects are used in the model, because in this situation the interest is in the specific lines. Reciprocal crosses are treated as distinct crosses. The differences in population mean and variance between reciprocals are assumed to be an indication for the inaccuracy of these estimates. However, we found some indications for reciprocal effects in a QTL-analysis of crosses GEI-119 × Gunhild and Gunhild × GEI-119 (Koorevaar, unpublished). This is an indication that the cytoplasm should be used as an additional factor in a QTL-analysis.

On the basis of the least squares means for the different traits (plant height, flowering time, lodging, thousand kernel weight, and grain yield), a segregation analysis is performed to

investigate the effect of 'major genes'. A robust mixture model (McLachlan and Basford, 1988) is fitted, using the idea for an iterative EM-algorithm described by Jansen (1993). It appeared possible to postulate 0 to 4 segregating 'major genes' per cross. Among them we have found the expected segregations of the *denso* gene (Haahr & Von Wettstein, 1976) and the *ert-g* gene (Thomas *et al.*, 1984). Segregation ratios appear to be significantly distorted for 25 of the 43 postulated 'major genes' (see chapter 3). One of these genes with distorted segregation is the *ert-g* gene. However, postulated 'major genes' may contain some inaccuracy. Using AFLP marker information for the RILs of the reciprocal crosses Apex × Prisma and Prisma × Apex (Yin, unpublished) we have not been able to find a marker associated with a postulated 'major gene' in that cross. Several candidate causes can be mentioned for this result. Maybe none of the available markers was tightly linked to the 'major gene'. Otherwise, the 'major gene' is not completely correctly postulated which may be due to the inaccuracy of the input data for the individual lines, the unjust normality assumptions in the mixture model, even when it is robust, and interaction between different genes, i.e., epistasis, causing segregating ratios deviating from the expected 1:1 ratio. In the marker analysis of the SSD-lines of both pairs of reciprocal crosses genome regions with distorted segregation have been found (Yin, Koorevaar, unpublished). This distorted segregation is also known from DH-line populations and it has been found for different regions of the genome (Devaux et al., 1996).

Mean prediction

Early-generation based prediction

The prediction of mean offspring performance is investigated in chapters 3 and 6. Most attention has been given to grain yield as it is genetically complex and very relevant for practical barley breeding. An early generation (F4) assessment of mean yield per cross appears not useful as a predictor of mean yield of a resulting recombinant inbred line (RIL) population. This is probably due to intergenotypic competition between F4-plots, mainly caused by the small plot size and the large variation in plant height. Usually it is not possible to extend the plot size in an early-generation trial because of lack of seed. The variation in plant height might be somewhat smaller in a practical breeding situation, though still present. In combination with the large environmental difference between small and large plots an early generation assessment will not allow a reliable prediction of mean offspring yield in barley under Dutch conditions. This is in agreement with the results in spring wheat presented by Van Ooijen (1989b).

Midparent-value based prediction

A well-known and simple alternative is also investigated, a prediction on the basis of midparent values, i.e., the average performance of the parents of the candidate cross. This prediction has appeared useful, especially for average yield over environments or thousand kernel weight. Due to genotype by environment interaction, a midparent value for yield based on one location-by-year combination is not a reliable predictor of average RIL yield over environments.

Significant variation for the difference between midparent yield and mean offspring yield has been observed. In most crosses midparent yield appears to be clearly higher than mean offspring yield. Thus, in these cases prediction on the basis of midparent value would result in an overestimation of the expected progeny yield. We have not found a way to predict in which crosses such an overestimation would occur and in which it would not. As an explanation for the difference between midparent yield and mean offspring yield we hypothesised distorted segregation and/or epistasis. Using molecular marker information, we have found some evidence for both hypotheses, but more investigation is necessary to draw reliable conclusions. In the case of distorted segregation towards the allele with a negative effect on yield, a reconsideration of the applied single seed descent method in comparison with other procedures to obtain homozygosity, e.g. doubled haploid systems, may be necessary. In the case of epistasis, when offspring segregation results in the loss of favourable parental allele combinations, it would be interesting to find ways to predict the degree of epistasis in candidate crosses. A QTL-analysis may be very useful in this aspect. The variation in the difference between midparent value and progeny mean is not found for thousand kernel weight and midparent values predict RIL means correctly.

Prediction of mean stability

Although midparent yields from a single environment are poor predictors of mean offspring yield averaged over environments, they are good predictors of mean offspring yield in the considered environment. The stability measures b_i (Finlay and Wilkinson, 1963), d_i^2 (Eberhart and Russell, 1966), and σ_i^2 (Shukla, 1972), investigated in chapter 6 are based on these parent and offspring yields and capture the genotype by environment interaction. The significant positive correlation between midparent values and offspring mean for b_i and d_i^2 can be explained by the strong correlation between the parent and offspring yields in a single environment. The stability variance σ_i^2 consists of two terms: one is based on b_i, the other is based on d_i^2. The lack of correlation between midparent value and progeny mean for σ_i^2 is a result of a difference between parents and offspring in the relative contribution of the two terms to σ_i^2. Although b_i and d_i^2 are both heritable, the opportunities for a reliable prediction of offspring mean for b_i are

expected to be much better than for d_i^2. This is based on the difference between b_i and d_i^2 in the strength of correlation between midparent value and offspring mean and on the difference in approximate standard errors for the two statistics. The relatively high standard error for d_i^2 is an indication of poor repeatability, i.e., the ranking of genotypes based on d_i^2 is strongly influenced by the investigated sample of environments. It is therefore concluded that from the three investigated stability statistics only b_i is predictable in a practical breeding situation. This is in agreement with the conclusion of Lin and Binns (1991), although they do not consider repeatability.

Variance prediction

Early-generation based prediction

The prediction of progeny variance is considered in chapters 3 to 6. A prediction of yield variance based on early generation trials is presented in chapter 3. Although CVs of the F4-trials are rather high and the genetic variance estimates are quite inaccurate, for the F4-lines as well as the RILs, the correlation coefficient between F4 variance and RIL variance is moderately high (r=0.62). However, this high correlation is mainly a result of differences between the crosses with respect to the number of segregating 'major genes'. These segregating 'major genes', e.g. the *denso* gene (Haahr & Von Wettstein, 1976) and the *ert-g* gene (Thomas *et al.*, 1984), were indentified in a segregation analysis and explain a large part of the yield variation within the RIL populations.

By elimination of 'major gene' effects, 'minor gene' variance is estimated. In this way we investigated the prediction of genetic variance for populations where no 'major genes' are segregating. This is a common situation in practical barley breeding. It may be a result of crosses between parents that both contain the favourable alleles of the 'major genes'. It can also be the result of visual selection for the favourable alleles in an early generation. So, the prediction of the 'minor gene' variance is very relevant in practice as it reveals the perspectives for selection in a normal offspring population. The 'minor gene' variance for yield, between F4-lines, shows a poor correlation with 'minor gene' variance for yield, between RILs. It is therefore concluded that a prediction of progeny variance for grain yield using early generation trials is not useful in practical barley breeding.

AFLP-based genetic similarity

In chapter 2 a genetic similarity, *gs*, based on AFLP-markers (Vos et al., 1995) is presented. gs_{ij} between genotype i and j is calculated following Nei and Li (1979), ignoring bands that are absent in both genotypes i and j. All bands from the investigated primer combinations are used,

including the monomorphic ones. This saves the 'existing' *gs* estimates from being recalculated every time marker information for a new genotype is added using the same primer combinations. Redundant bands within primer combinations are discarded because they are usually derived from the same genome position (Qi and Lindhout, 1997). The application of AFLPs for barley cultivar identification is useful because for each of primer combinations we used all investigated barley genotypes showed distinguishable marker patterns. AFLPs also seem applicable for the assessment of genetic diversity, as a set of barley genotypes is easily divided into major ecotypes on the basis of *gs* estimates. A third application is the prediction of progeny variance, which is discussed in the next paragraph.

Prediction based on parental relationship measures

The genetic variance within an offspring population can be predicted by parental relationship measures. This prediction is based on the idea, described in chapter 1, that differences in genetic variance between crosses are mainly a result of differences in the number of segregating genes. This number of segregating genes is expected to be negatively correlated with the relatedness of the parents. However, the usefulness of parental relationship measures for direct variance prediction depends largely on the agreement between the genomic representation caught by the parental relationship measure and the genomic distribution of the segregating genes that make up the genetic variance among the offspring.

Because pedigree information is lacking for part of the parents, only 10 crosses are used to examine the predictive value of the coefficient of coancestry f (Malécot, 1948). Although correlations between $1-f$ and offspring variances are mainly positive, the number of crosses is too small to draw reliable conclusions. However, pedigree data can naturally be considered as a good source of information for close relationships.

A distance measure, *md*, based on parent data for 25 morphological traits (UPOV, 1981) does not appear to be a good predictor of progeny variance. This may be due to the poor accuracy of this distance measure, caused by the rather rough ordinal scales that are used to score most of the traits. Burstin and Charcosset (1997) show that similarities in the phenotypes of two parents do not have to be a result of genetic relatedness. It is also questionable whether the positions of the genes underlying the morphological traits are representative for the positions of genes underlying the agronomic traits for which progeny variance is considered.

The latter problem is more or less solved by the use of the agronomic distance measure, *agd*, based on the parental values for the investigated agronomic trait. Poor-to-moderate correlations are found between these trait specific *agd*s and trait variance, except for yield, for which the correlation is absent. However, the correlations appear to be based on 'major gene' effects. An allele difference between the parents for a 'major gene' causes a large trait value difference

between parents as well as a large variance among the offspring. This effect is clearly observed for plant height and flowering time as a result of absence or presence of the mutant-allele of the *denso*-gene, and for plant height and thousand kernel weight considering the *ert-g* gene. The poor and non-significant correlations between *agd* and progeny variance for yield and between *agd* and 'minor gene' variances corroborate conclusions of Burstin and Charcosset (1997). They state that large phenotypic differences usually agree with large genotypic differences. However, small phenotypic differences are not necessarily based on small genotypic differences.

In the case of yield a clear genotype by environment interaction was observed. This information is not effectively used in the *agd*-based prediction of progeny variance for yield, because the agronomic distance is largely determined by differences in the general yield level of the parents. Large differences between parents for yield in specific environments have a relatively small impact. One can focus on these differences by examining the residuals of an analysis of variance using a model with additive genotype and environment effects. The correlation coefficient r_{GxE} between the environment-specific responses of two parents, i.e., the residuals, appears to be a relatively good predictor of yield variance among the offspring. This confirms earlier results of Habgood (1983), who predicted F2-variance for several yield components in barley. On the basis of this correlation between r_{GxE} and progeny variance for yield, we suppose that there is a positive correlation between the overall effect of a yield QTL and the QTL by environment interaction variance of that QTL. It seems that the majority of yield QTL are sensitive to changes in environments. The usefulness of the correlation coefficient between parental environment responses as a predictor for progeny yield variances in a practical breeding programme may be dependent on 'major gene' effects. Because the effect of the 'major genes' on the parental environment-specific responses cannot be estimated in this study, no final conclusions can be drawn. Further investigation using populations without segregating 'major genes' is probably useful.

The correlation coefficients between AFLP-based genetic distance ($1-gs$) and progeny variances are generally positive, but non-significant. Analogously to the lack of correlation between genetic distance and heterosis observed by Charcosset et al. (1991), the lack of correlation between marker-based genetic distance and trait variance is most likely a result of markers that are not linked to QTL for the investigated trait and QTL that are not linked to markers included in *gs* estimation. One could thus say that random markers are a poor representation of the segregating trait genes and their effects.

In an attempt to improve the genomic representation of markers, genetic map information was used. The resulting map-based genetic distance, *mgd*, is more or less independent of the marker density on the map. The correlation of 'minor gene' offspring variance with *mgd* is just as poor as with $1-gs$. The assumption that 'minor gene' variance is based on many segregating genes with small effects, referring to the general quantitative genetic model (Mather and Jinks,

1982), is probably not completely fulfilled, suggesting that mionor genes and their effects are not uniformly scattered over the genome in large numbers. Another reason for the poor predictive value of *mgd* is the lack of representation of the expressed part of the genome ('the genes') by the markers of the recombination map: the weighting of markers that we applied does of course not remove gaps in the linkage map.

In order to estimate $1-gs$ with AFLP-markers representative for those parts of the genome which are primarily responsible for the genetic variance, marker selection was performed. Markers are selected on the basis of a strong marker-trait association in the parent population. The degree of association is based on an F-test for the marker effect in an analysis of variance of the investigated trait using data from several environments. The genetic distance, *sgd*, is based on the selected markers. To obtain a good and reliable predictor *sgd* for progeny variance for the examined trait, several marker selection procedures and selection thresholds are compared. To include part of the possible epistatic effects and to prevent selection of markers that are loosely linked to 'major genes', a marker selection procedure was applied that uses a model with an extra marker additional to the marker under investigation. Each of the available markers, except the investigated one, is used as extra marker in a separate model. The P-values for the investigated marker are averaged over all these 'augmented' models. Markers selected by using this average P-value as a selection criterion produce *sgd* estimates that give higher correlations with progeny variance than *sgd*s based on the P-value of a single-marker model or $1-gs$ based on all available markers. Without knowing the map position of the markers, it is difficult to get insight in the exact action of the 'extra marker' model while selecting markers. It is also not clear how to establish a generally applicable selection threshold for the average P-value. Selection intensity may be a better criterion than an absolute selection threshold, although this intensity may still vary depending on crop and trait. In order to account for the effect of the QTL that the selected marker is expected to be associated with, the fraction of the trait variance explained by that marker is used as a weight in the *sgd* calculation. In general the correlation coefficient of progeny variance with this weighted *sgd* is not higher than with an unweighted *sgd*. We may conclude that marker-trait associations in a parent population of only 18 parents are informative enough to select markers that represent genomic regions responsible for the genetic variance of a trait. The estimation of the relative contribution of these regions is, however, not reliable enough to be useful in *sgd* assessment. The effectiveness of the marker selection procedure may be improved by sampling more parents from the same group of rather closely related European two-row spring barleys. The observed correlations between *sgd* and progeny variance in this study, in situations with and without segregating 'major genes', are not high enough to allow reliable variance prediction in practical plant breeding.

Another approach to combine relatedness information from different sources is a direct pooling of several distance estimates. In general, the investigated relationship measures based

on different types of information show no significant correlation with each other, if tested under the correct assumptions. It is quite conceivable that they represent different parts of the genome. They can also be quite inaccurate. The independent inaccuracies may be decreased by pooling the various distance estimates (Cox et al., 1985). The first component from a principal components analysis of the distance estimates $1-f$, $1-gs$, *md,* and *agd* generally shows higher correlation coefficients with progeny variance than the separate distance estimates. The correlation coefficients between the following pooled distance measure and offspring variance seem to be even higher. This pooled distance is based on $1-f$ and $1-gs$ for closely related parent combinations and on *agd* and *md* for distantly related parent combinations. These higher correlations show that it may be useful to employ those distance estimates that are expected to be the most reliable in a certain parent combination. However, the division between distantly and closely related parent combinations is rather arbitrary. It can be concluded that this type of combined distance measure needs more investigation. Distance estimates based on marker selection and correlation between environment-specific responses of genotypes can be included.

An interesting application of parental relationship measures is presented by Bernardo (1994) and Charcosset et al. (1998) and already discussed in chapter 1. They use genetic distances to describe the relationship between tested predictor hybrids and candidate hybrids, i.e., potential parent combinations. Although it has not been investigated in this study, it may be possible to predict mean and variance of candidate crosses in a similar way. A limited set of crosses is made and DH or SSD lines are tested to obtain mean and variance estimates for the predictor crosses. Best linear unbiased predictions (Bernardo, 1994) or factorial regression estimators (Charcosset et al., 1998) can be used to predict mean and variance of candidate crosses. Finally, it may also be interesting to test the usefulness of the parental relationship measures, introduced in the present study, for hybrid prediction, following Bernardo (1994) or Charcosset et al. (1998).

Cross prediction combining mean and variance

The predicted mean and variance can be combined, following Jinks and Pooni (1976), to obtain a prediction of the probability that an inbred line descending from the investigated parent combination performs better than an arbitrary threshold level. In chapter 3 this procedure is tested for grain yield with a mean prediction based on midparent values and a variance prediction based on the variance among F4-lines. The correlations between predictor and observed value are 0.71 for progeny mean and 0.62 for progeny variance. However, the combination of mean and variance prediction resulted in a very poor prediction of the number of RILs exceeding the threshold (r_s=0.22). This is basically caused by three factors: (1) the lack

of predictive value of the midparent value, probably due to distorted segregation and epistasis, (2) the lack of predictive value of the F4 variances, probably due to inaccurate variance estimates, genotype by environment interaction and interplot competition, (3) the sampling error of the number of RILs exceeding the threshold value. Also the normality assumptions in the prediction method may have been inappropriate because of 'major gene' effects. However, a prediction using 'major gene' information, i.e., segregation ratios and allele effects, hardly improves the correlation coefficient between the combined cross prediction and the observed number of desired SSD-lines. Earlier studies in small cereals (Snape, 1982; Tapsell and Thomas, 1983) mention reasonable results of cross prediction based on predicted mean and variance. However, all material was grown in the same environment, excluding the effect of genotype by environment interaction, and the effect of segregating 'major genes' was not mentioned. If both crosses with and without segregating 'major genes' were used, results are indeed expected to be better than in the situation without segregating 'major genes'.

Line prediction

Within crosses, the yields of the F4-lines are compared to the yields of the RILs descending from the same F2-plant. A strong correlation between F4-yield and RIL-yield would open perspectives for early generation selection for yield. However, the observed correlation is poor and slightly negative. It appears to be based on segregating 'major genes'. Such genes can easily be selected by visual observation, which is more cost-effective than a laborious yield assessment. We find evidence for interplot competition in the small F4 plots based on a positive correlation between plant height and yield. This correlation between plant height and yield becomes slightly negative when considering the large plot RIL trials.

Cross prediction in practice

Plant breeding programmes generally include cross prediction, explicitly or implicitly, as described in chapter 1. The present study demonstrates the usefulness of some of the explicit cross prediction methods. The usefulness may depend on crop and/or trait, so some additional examination will be necessary to extend the results from this barley study to other breeding programmes. However, most of the cross prediction methods that are presented can easily be applied in other selffertilizing crops. Mean and variance prediction methods can be used to predict inbred line performance in hybrid cultivar breeding as well as pure line cultivar breeding. In crops like maize, cabbage and tomato, the performance of the inbred line is judged by its performance as a hybrid parent. In barley, as well as in crops like wheat and lettuce, the

performance of the inbred line is judged by its performance as a cultivar. Moreover, the prediction of the performance of hybrid cultivars is often divided in prediction of general combining ability and prediction of heterosis, which can be regarded as parallels of mean and variance prediction.

A prediction of mean offspring performance on the basis of midparent values is often applied in practice. It can be a good criterion for the choice of parent combinations, also when considering more than one trait. The results of the present study support a mean prediction based on midparent values. However, for grain yield we observed significant deviations from the predicted means. Whatever the cause, distorted segregation or epistasis, a breeder should be aware of the possible occurence of these deviations and, if they are large, try to predict their size for the different parent combinations. The assessment of parent traits is usually cheap compared to the costs of making the candidate crosses and evaluating offspring performance. It is obvious that candidate parents should be evaluated in more than one environment for traits that show genotype by environment interaction.

An explicit prediction of offspring variance is not often performed in practical plant breeding. However, breeders are anxious to obtain transgressive segregants, i.e., descendants that perform better than the best parent. First, this can be achieved by using almost equally well performing genotypes as parent. If the 'best' parent is kept the same, this means increasing the mean performance of the offspring. Second, transgression can be further increased by using parent combinations that are expected to give a highly variable offspring population. The results of the present study show that the prediction of variance is difficult, especially when there are no 'major genes' segregating. If 'major genes' segregate, it is usually possible to predict the mean performance of the offspring after visual selection for the desired alleles. The predictors of offspring variance are mainly parental relationship measures. Considering the results of the present study, it may be useful to apply some of the parental relationship measures as indicators for progeny variance. The word 'indicator' is used instead of 'predictor' to stress the supporting role of these parental relationship measures in the decision which parent combinations will proceed to a crossing and selection program. On the basis of the indicators, a breeder could for instance decide whether to test a large or a small number of descendants from a parent combination. In a breeding program the requirements for an offspring line to become a cultivar often include certain minimum levels of performance for several traits at the same time. In this case it may sometimes be decided to discard a parent combination because of the unaccessibly large amount of variation that is expected. For such a parent combination the probability that an offspring line performs equal to or better than the minimum levels for all traits, is very small, suggesting that a proper offspring evaluation would require a disproportional amount of resources.

References

Arias G, Reiner A, Penger A, Mangstl A (1983) Directory of barley cultivars and lines. Ulmer Press, Stuttgart

Autrique E, Nachit MN, Monneveux P, Tanksley SD, Sorrells ME (1996) Genetic diversity in durum wheat based on RFLPs, morphological traits and coefficient of parentage. Crop Sci 36:735-742

Barbosa-Neto JF, Sorrells ME, Cisar G (1996) Prediction of heterosis in wheat using coefficient of parentage and RFLP-based estimates of genetic relationship. Genome 39:1142-1149

Barnett V, Lewis T (1994) Outliers in statistical data. Third edition. John Wiley and Sons, Chichester. 584 pp.

Baum BR, Bailey LG, Thompson BK (1985) Barley register. Agric Canada Publ 1783B. Ottawa, Canada

Becker HC, Léon J (1988) Stability analysis in plant breeding. Plant Breeding 101:1-23

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20-25

Boppenmaier J, Melchinger AE, Seitz G, Geiger HH, Herrmann RG (1993) Genetic diversity for RFLPs in european maize inbreds. III. Performance of crosses within versus between groups for grain traits. Plant Breeding 111:217-226

Bos I, Caligari PDS (1995) Selection methods in plant breeding. Chapman and Hall, London

Bos I, Sparnaaij LD (1993) Component analysis of complex characters in plant breeding. II. The pursuit of heterosis. Euphytica 70:237-245

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

Briggs KG, Shebeski LH (1971) Early generation selection for yield and breadmaking quality of hard red spring wheat (*Triticum aestivum* L. Em Thell.). Euphytica 20:453-463

Burkhamer RL, Lanning SP, Martens RJ, Martin JM, Talbert LE (1998) Predicting progeny variance from parental divergence in hard red spring wheat. Crop Sci 38:243-248

Burstin J, Charcosset A (1997) Relationship between phenotypic and marker distances: theoretical and experimental investigations. Heredity 79:477-483

Caligari PDS, Brown J (1986) The use of univariate cross prediction methods in the breeding of a clonally reproduced crop (Solanum tuberosum). Heredity 57:395-401

Caligari PDS, Powell W, Jinks JL (1985) The use of doubled haploids in barley breeding. 2. An assessment of univariate cross prediction methods. Heredity 54:353-358

Charcosset A, Essioux L (1994) The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. Theor Appl Genet 89:336-343

Charcosset A, Lefort-Buson M, Gallais A (1991) Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. Theor Appl Genet 81:571-575

Charcosset A, Bonnisseau B, Touchebeuf O, Burstin J, Dubreuil P, Barriere Y, Gallais A, Denis JB (1998) Prediction of maize hybrid silage performance using marker data: Comparison of several models for specific combining ability. Crop Sci 38: 38-44

Choo TM (1988) Cross prediction in barley using doubled-haploid lines. Genome 30:366-371

Cowen NM, Frey KJ (1987a) Relationship between genealogical distance and breeding behaviour in oats (*Avena sativa* L.). Euphytica 36:413-424

Cowen NM, Frey KJ (1987b) Relationships between three measures of genetic distance and breeding behavior in oats (*Avena sativa* L.). Genome 29:97-106

Cox TS, Murphy JP (1990) The effect of parental divergence on F2 heterosis in winter wheat crosses. Theor Appl Genet 79:241-250

Cox TS, Kiang YT, Gorman MB, Rodgers DM (1985a) Relationship between coefficient of parentage and genetic similarity indices in the soybean. Crop Sci 25:529-532

Cox TS, Lookhart GL, Walker DE, Harrell LG, Albers LD, Rodgers DM (1985b) Genetic relationships among hard red winter wheat cultivars as evaluated by pedigree analysis and gliadin polyacrylamide gel electrophoretic patterns. Crop Sci 25:1058-1063

DePauw RM, Shebeski LH (1973) An evaluation of an early generation yield testing procedure in *Triticum aestivum*. Can J Plant Sci 53:465-470

Devaux P, Zivy M, Kilian A, Kleinhofs A (1996) Doubled haploids in barley. In: Scoles JG, Rossnagel BG (eds) V International Oat Conference & VII International Barley Genetics Symposium: Proceedings. volume I. Univ of Saskatchewan, Saskatoon. pp 213-222

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297-302

Dillmann C, Charcosset A, Goffinet B, Smith JSC, DattéeY (1997) Best linear unbiased estimator of the molecular genetic distance between inbred lines. In: Krajewski P, Kaczmarek Z (eds) Advances in biometrical genetics. Proc. of the 10th meeting Eucarpia section Biometrics in Plant Breeding. Poznań, 14-16 May 1997, pp 105-110

Dudley JW, Saghai-Maroof MA, Rufener GK (1991) Molecular markers and grouping of parents in maize breeding programs. Crop Sci 31:718-723

Eberhart SA, Russell WA (1966) Stability parameters for comparing varieties. Crop Sci 6:36-40

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant-breeding programme. Aust J Agric Res 14:742-754

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

Gollob HF (1968) A statistical model which combines features of factor analytic and analysis of variance techniques. Psychom 33:73-115

Goodman MM (1972) Distance analysis in biology. Syst Zool 21:174-186

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325-338

Graner A, Ludwig WF, Melchinger AE (1994) Relationships among European barley germplasm: II. Comparison of RFLP and pedigree data. Crop Sci 34:1199-1205

Haahr V, Von Wettstein D (1976) Studies of an induced high yielding dwarf-mutant of spring barley. In: Proc 3rd Int Barley Genet Symp, Garching. 215-218

Habgood RM (1977) Estimation of genetic diversity of self-fertilizing cereal cultivars based on genotype-environment interactions. Euphytica 26:485-489

Habgood RM (1983) Genotype-environment interactions for predicting the breeding value of bi-parental crosses in spring barley. Euphytica 32:273-279

Hayes PM, Liu BH, Knapp SJ, Chen F, Jones B, Blake T, Franckowiak J, Rasmusson D, Sorrells M, Ullrich SE, Wesenberg D, Kleinhofs A (1993) Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. Theor Appl Genet 87:392-401

Helms T, Orf J, Vallad G, McClean P (1997) Genetic variance, coefficient of parentage, and genetic distance of six soybean populations. Theor Appl Genet 94:20-26

Hockett EA, Cook AF, Khan MA, Martin JM, Jones BL (1993) Hybrid performance and combining ability for yield and malt quality in a diallel cross of barley. Crop Sci 33:1239-1244

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Statist 35:73-101

Jansen RC (1992) A general mixture model for mapping quantitative trait loci by using molecular markers. Theor Appl Genet 85:252-260

Jansen RC (1993) Maximum likelihood in a generalised linear finite mixture model by using the EM-algorithm. Biometrics 49:227-231

Jinks JL, Pooni HS (1976) Predicting the properties of recombinant inbred lines derived by single seed descent. Heredity 36:253-266

Jinks JL, Pooni HS (1980) Comparing predictions of mean performance and environmental sensitivity of recombinant inbred lines based upon F3 and triple test cross families. Heredity 45:305-312

Kearsey NJ, Jinks JL (1968) A general method of detecting additive, dominance and epistatic components of variation for a metrical trait. I. Theory. Heredity 23:403-409

Kempthorne O, Curnow RN (1961) The partial diallel cross. Biometrics 17:229-250

Knott DR (1979) Selection for yield in wheat breeding. Euphytica 28:37-40

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.

Li Z, Pinson SRM, Park WD, Paterson AH, Stansel JW (1997) Epistasis for three grain yield components in rice (*Oryza sativa* L.). Genetics 145:453-465

Lin CS, Binns MR (1988) A method for analyzing cultivar x location x year experiments: a new stability parameter. Theor Appl Genet 76:425-430

Lin CS, Binns MR (1991) Genetic properties of four types of stability parameter. Theor Appl Genet 82:505-509

Lin CS, Binns MR, Lefkovitch LP (1986) Stability analysis: where do we stand? Crop Sci 26:894-900

Loiselle F, Tai GCC, Christie BR (1991) Pedigree, agronomic and molecular divergence of parents in relation to progeny performance in potato. Potato Research 34:305-316

Mahalanobis PC (1936) On the generalized distance in statistics. Proc Nat Inst Sci India 2:49-55

Malécot G (1948) Les mathématiques de l'hérédité. Masson & Cie, Paris

Manjarrez-Sandoval P, Carter TE Jr., Webb DM, Burton JW (1997) RFLP genetic similarity estimates and coefficient of parentage as genetic variance predictors for soybean yield. Crop Sci 37:698-703

Maronna RA (1976) Robust M-estimators of multivariate location and scatter. Ann Statist 4:51-67

Martin JM, Talbert LE, Lanning SP, Blake NK (1995) Hybrid performance in wheat as related to parental diversity. Crop Sci 35:104-108

Mather K, Jinks JL (1982) Biometrical genetics. 3rd edition. Chapman and Hall, London

Mather K, Vines A (1952) The inheritance of height and flowering time in a cross of *Nicotiana rustica*. In: Reeve ECR, Waddington CH (eds) Quantitative Inheritance. H.M.S.O., London, pp 49-80.

McCullagh P, Nelder JA (1989) Generalized linear models. Second edition. Chapman and Hall, London

McLachlan GJ, Basford KE (1988) Mixture models. Inference and applications to clustering. Marcel Dekker, New York

Melchinger AE, Graner A, Singh M, Messmer MM (1994) Relationships among European barley germplasm: I. Genetic diversity among winter and spring cultivars revealed by RFLPs. Crop Sci 34:1191-1199

Messmer MM, Melchinger AE, Herrmann RG, Boppenmaier J (1993) Relationship among early European maize inbreds: II. Comparison of pedigree and RFLP data. Crop Sci 33:944-950

Moser H, Lee M (1994) RFLP variation and genealogical distance, multivariate distance, heterosis, and genetic variance in oats. Theor Appl Genet 87:947-956

Neele AEF (1990) Study on the inheritance of potato tuber yield by means of harvest index components and its consequences for choice of parental material. Euphytica 48:159-166

Neele AEF, Nab HJ, Louwes KM (1991) Identification of superior parents in a potato breeding programme. Theor Appl Genet 82:264-272

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269-5273

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding 2:225-238

Qi X, Lindhout WH (1997) Development of AFLP markers in barley. Mol Gen Genet 254:330-336

Qi X, Stam P, Lindhout WH (1998) Use of locus-specific AFLP markers to construct a high-density moleculare map in barley. Theor Appl Genet 96:376-384

Rousseeuw PJ, Van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. J Amer Statist Assoc 85:633-639.

SAS Institute Inc (1989) SAS/IML® software: usage and reference, version 6, first edition. SAS Institute Inc, Cary, NC.

SAS Institute Inc. (1992) SAS technical report P-229, SAS/STAT® software: changes and enhancements, release 6.07. SAS Institute Inc., Cary, NC

Schut JW (1997) Application of the bootstrap in plant genetics. In: Stein A, Penning de Vries FWT, Schut JW (eds) Data in action. Proceedings of a seminar series 1996/1997. DLO Research Institute for Agrobiology and Soil Fertility and C.T. de Wit Graduate School for Production Ecology, Wageningen (Quantitative Approaches in Systems Analysis No. 12, pp 71-79)

Shukla GK (1972) Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237-245

Smith OS, Smith JSC, Bowen SL, Tenborg RA, Wall SJ (1990) Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, grain yield, heterosis, and RFLPs. Theor Appl Genet 80:833-840

Snape JW (1982) Predicting the frequencies of transgressive segregants for yield and yield components in wheat. Theor Appl Genet 62:127-134

Sneep J (1977) Selection for yield in early generations of self-fertilizing crops. Euphytica 26:27-30

Sneller CH, Kilgore-Norquest L, Dombek D (1997) Repeatability of yield stability statistics in soybean. Crop Sci 37:383-390

Sokal RR (1965) Statistical methods in systematics. Biol Rev Camb Phil Soc 40: 337-341

Souza E, Sorrells ME (1991a) Relationships among 70 North American oat germplasms: I. Cluster analysis using quantitative characters. Crop Sci 31:599-605

Souza E, Sorrells ME (1991b) Relationships among 70 North American oat germplasms: II. Cluster analysis using qualitative characters. Crop Sci 31:605-612

Souza E, Sorrells ME (1991c) Prediction of progeny variation in oat from parental genetic relationships. Theor Appl Genet 82:233-241

Spitters CJT (1979) Competition and its consequences for selection in barley breeding. Agr Res Report 893. Pudoc, Wageningen

Stuber CW (1992) Biochemical and molecular markers in plant breeding. In: Janick J (ed) Plant Breeding Reviews 9. John Wiley & Sons, New York. pp 37-61

Tapsell CR, Thomas WTB (1983) Cross prediction studies on spring barley. 2. Estimation of genetical and environmental control of yield and its component characters. Theor Appl Genet 64:353-358

Thomas WTB (1987) The use of random F3 families for cross prediction in spring barley. J Agric Sci, Camb 108:431-436

Thomas WTB, Powell W, Wood W (1984) The chromosomal location of the dwarfing gene present in the spring barley variety Golden Promise. Heredity 53:177-183

Tinker NA, Fortin MG, Mather DE (1993) Random amplified polymorphic DNA and pedigree relationships in spring barley. Theor Appl Genet 85:976-984

Tinker NA, Mather DE, Rossnagel BG, Kasha KJ, Kleinhofs A, Hayes PM, Falk DE, Ferguson T, Shugar LP, Legge WG, Irvine RB, Choo TM, Briggs KG, Ullrich SE, Franckowiak JD, Blake TK, Graf RJ, Dofing SM, Maroof MAS, Scoles GJ, Hoffman D, Dahleen LS,

Kilian A, Chen F, Biyashev RM, Kudrna DA, Steffenson BJ (1996) Regions of the genome that affect agronomic performance in two-row barley. Crop Sci 36:1053-1062

Tukey JW (1958) Bias and confidence in not quite large samples. Ann Math Statist 29:614

UPOV (1981) Guidelines for the conduct of tests for distinctness, homogeneity and stability: barley. UPOV, Geneva

Van Berloo R, Stam P (1998) Comparison between marker assisted selection and phenotypical selection in a set of Arabidopsis thaliana recombinant inbred lines. Theor Appl Genet (in press)

Van der Beek JG, Verkerk R, Zabel P, Lindhout WH (1993) Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: *Cf9* (resistance to *Cladosporium fulvum*) on chromosome 1. Theor Appl Genet 84:106-112

Van Eck HJ, Van der Voort JR, Draaistra J, Van Zandvoort P, Van Enckevort E, Segers B, Peleman J, Jacobsen E, Helder J, Bakker J (1995) The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. Molecular Breeding 1:397-410

Van Eeuwijk, FA (1995) Linear and bilinear models for the analysis of multi-environment trials: I An inventory of models. Euphytica 84:1-7

Van Hintum TJL, Haalman D (1994) Pedigree analysis for composing a core collection of modern cultivars, with examples from barley (*Hordeum vulgare* s. lat.). Theor Appl Genet 88:70-74

Van Oeveren AJ (1992) A comparison between single seed descent and early cross selection in wheat breeding. Euphytica 58:275-287

Van Ooijen JW (1989a) The predictive value of quantitative genetic parameters in autogamous crops: bias caused by intergenotypic competition. 1. F∞-mean. Euphytica 44:81-94

Van Ooijen JW (1989b) The predictive value of quantitative genetic parameters in autogamous crops: bias caused by intergenotypic competition. 2. F∞-variance. Euphytica 44:95-108

Virk PS, Ford-Lloyd BV, Jackson MT, Pooni HS, Clemeno TP, Newbury HJ (1996) Predicting quantitative variation within rice germplasm using molecular markers. Heredity 76:296-304

Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M, (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23: 4407-4414

Wilks SS (1963) Multivariate statistical outliers. Sankhya Ser A 25:407-426

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531-6535

Zadoks JC, Chang TT, Konzak CF (1974) A decimal code for the growing stages of cereals. Eucarpia Bull 7:42-52

Zhang QF, Gao, YJ, Yang SH, Ragab RA, Maroof MAS, Li ZB (1994) A diallel analysis of heterosis in elite hybrid rice based on RFLPs and microsatellites. Theor Appl Genet 89:185-192

Zobel RW, Wright MJ, Gauch HG (1988) Statistical analysis of a yield trial. Agron J 80:388-393

Summary

In plant breeding programmes genetic variation has to be created in order to be able to select new cultivars. This variation is often created by crossing genetically divergent parents. The choice of parent combinations determines the genetic variation on which new cultivars will be based and which genes will be (re)combined by crossing. To provide a solid basis for the choice of parent combinations the performance of the offspring can be predicted using knowledge about the genes underlying important traits and the corresponding parental genotypes. If this knowledge is absent or incomplete, other prediction methods can be used. These are largely based on the relationship between parents and offspring, either inferred empirically or derived from genetic theory.

The subject of the study is the prediction of offspring mean and variance in an inbred crop, using the latter type of prediction methods. Several approaches are compared and investigated for their usefulness in practical plant breeding. One of the sources of parent information used for cross prediction, is a relatively new type of molecular markers: AFLPs. European two-row spring barley is used as a model crop. To obtain reliable results and draw general conclusions, 20 crosses, each represented by 48 recombinant inbred lines (RILs), are tested with their parents in large 10-row plots in 7 environments, distributed over two years. Offspring performance is observed for four agronomic traits: plant height, flowering time, thousand kernel weight and grain yield.

In chapter 2 a genetic similarity, *gs*, based on AFLP markers is compared with other parental relationship measures. The AFLP-based *gs* shows a poor-to-moderate correlation with the coefficient of coancestry, f, based on pedigree data. No correlation is found with morphological distance, *md*, based on data for 25 morphological traits. Bootstrap sampling from the parental genotypes is performed to assess the accuracy of the estimated correlation coefficients between the relationship measures. The AFLP-based genetic similarity appears useful to assess some of the major ecotypes of barley, and AFLPs, even if they are derived from only one primer combination, appear useful for cultivar identification.

In chapter 4 the usefulness of the parental relationship measures *gs*, f, and *md* for the prediction of progeny variance is examined. Correlations between $1-f$ and offspring variance are mainly positive, but non-significant. They are based on only 10 parent combinations which have reliable pedigree data. Correlations between *md* and variance among the RILs are non-significant. Correlations between $1-gs$ and progeny variance for the investigated traits are generally positive, but seldom significant. The poor correlations are expected to be a result of a poor genomic representation by AFLP markers of the genes affecting the traits. Another parental relationship measure is introduced: agronomic distance, *agd*, based on multi-environment data for

several agronomic traits. The correlation between *agd* and offspring variance is mainly positive, and sometimes significant, especially when the traits, used in the calculation of the parental *agd* and the RIL variance, are the same. However, this correlation appears to be due to differences between crosses in the number and effect of 'major genes'. These genes can be visually selected in early generations. The 'minor gene' variance remaining after visual selection for desired alleles of 'major genes', cannot reliably be predicted by *agd*. Combined relationship measures generally have the highest correlations with progeny variance. This is especially true when different combinations of relationship measures are used for closely and distantly related parent combinations. $1-gs$ and $1-f$ seem more reliable for related parent combinations, while *agd* and *md* may be more reliable for distant parent combinations. However, the correlations between the combined relationship measures and progeny variance are not high enough to allow a reliable variance prediction.

In chapter 5 two possible improvements of the genomic representation of investigated traits by AFLP markers are studied. First, the overrepresentation of genomic regions caused by clustered markers is eliminated by the use of genetic map information. Markers are weighted for marker density around their map position in the calculation of the genetic distance *mgd*. The correlation of progeny variance with *mgd* is just as poor as with $1-gs$ using unweighted markers. So, perspectives for *mgd*-based variance prediction are poor. Secondly, a genetic distance *sgd* is calculated using only those markers that are selected on the basis of strong marker-trait associations in the population of parents. The applied selection criterion is the P-value of the F-test for the investigated marker in an analysis of variance of the parent data. Several marker selection procedures and selection thresholds are tested. The correlation between *sgd* and progeny variance is highest when a range of modified ANOVA models is used for marker selection. These models each include the investigated marker plus one of the other available markers in the parental data set. The average P-value for the investigated marker in these models is used as selection criterion. For the different traits we found an optimum selection threshold for the average P-value of 0.005. Only for flowering time it appeared useful to include marker by environment interaction in the model. However, correlations between *sgd* and progeny variance are not high enough to be reliably used for variance prediction in practical breeding.

In chapter 3 we investigate the usefulness of an early generation (F4) small-plot trial over two environments to predict offspring mean and variance for grain yield. The mean yield of the RILs cannot be predicted by the mean yield of the F4-lines, because of interplot competition between the small plots and because of genotype by environment interaction. The midparent value for yield over environments, measured in large plots, appears to be a reasonable predictor of mean RIL yield over environments (r=0.71). However, for most crosses the midparent value overestimates the mean offspring yield, probably as a result of distorted segregation and epistasis. F4 variance for yield is moderately correlated with RIL variance (r=0.62). However, this

relationship is based on differences between crosses in the number and effects of segregating 'major genes'. The 'minor gene' variance remaining after visual selection for desired alleles of 'major genes', shows a poor correlation with F4 variance (r=0.41). The combination of predicted mean and variance results in a poor prediction of the probability for a line in the RIL population to exceed a certain threshold level. Selection for yield of F4-lines within a cross is also not useful in practice, because of interplot competion, a large standard error for individual line estimates and genotype by environment interaction.

In chapter 6 the prediction of offspring performance in the presence of genotype by environment interaction is examined. First, midparent values for three stability parameters are used to predict the average stability of the offspring populations. A reasonable correlation between midparent value and RIL mean is found for the coefficient of regression of parent or RIL performance on the environmental index, i.e., the average of all parents in a certain environment (r_s=0.69). A weaker, but significant correlation is found for the second measure, the mean squared deviation from the regression (r_s=0.39). The correlation is virtually absent for the third measure, the stability variance, because its composition is different for parents and RILs (r_s=0.17). Because repeatability of the mean squared deviation is much poorer than that of the regression coefficient, it is concluded that only a prediction of the average regression coefficient of an offspring population is useful in practice. The genotype by environment interaction of parents and offspring is further assessed with an AMMI-analysis in which the interaction is described by a parsimonious set of multiplicative parameters. Although there is a clear correlation between midparent values and offspring means for these parameters, perspectives for a prediction of the offspring interaction are limited. The interaction pattern of parents can also be used to assess the relationships between them. As a similarity measure we use the correlation coefficient $r_{G\times E}$. It is based on the correlation between the parental residual vectors from an analysis of variance of the multi-environment yield data for the parents. An ANOVA model with additive genotype and environment effects is used. The calculated $r_{G\times E}$ is negatively correlated with RIL variance for yield over environments (r_s=−0.63). We conclude that this relationship measure may be useful for variance prediction, although further investigation is necessary.

In conclusion, it is stated that mean offspring prediction on the basis of midparent values is useful in practical breeding. Prediction of progeny variance is less reliable, especially when no 'major genes' are segregating. It is proposed to use parental relationship measures as 'indicators' for progeny variance to stress their supporting role in the choice of parent combinations that will proceed to a crossing and selection program. Several investigated relationship measures are useful to make a rough distinction between uniform and variable offspring populations. The degree of usefulness, although crop and trait dependent, may be approximately indicated by the strength of the correlation of the relatedness measure with progeny variance found in this study.

Samenvatting

In een plantenveredelingsprogramma moet eerst genetische variatie gecreëerd worden om daarna nieuwe rassen te kunnen selecteren. Vaak wordt deze variatie tot stand gebracht door het kruisen van genetisch verschillende ouders. De keuze van de oudercombinaties voor het maken van kruisingen bepaalt welk deel van de beschikbare genetische variatie benut wordt en welke genen ge(re)combineerd worden in de nakomelingen. Ter ondersteuning van de keuze van oudercombinaties kan men de prestatie van de nakomelingen voorspellen met behulp van kennis over de genen voor een eigenschap en hun bijbehorende oudergenotypen. Als deze kennis geheel of gedeeltelijk ontbreekt, kunnen andere voorspellingsmethoden worden gebruikt. Deze methoden maken voornamelijk gebruik van relaties tussen ouders en nakomelingen, die ofwel empirisch bepaald zijn ofwel berusten op genetische theorie.

Het onderwerp van dit proefschrift is het gebruik van de laatstgenoemde voorspellingsmethoden als basis voor de keuze van oudercombinaties in een zelfbevruchtend gewas. In dit gewas worden het gemiddelde en de variantie voor de nakomelingschap van een oudercombinatie voorspeld. Verschillende methoden zijn met elkaar vergeleken en onderzocht op hun bruikbaarheid voor voorspelling in de veredelingspraktijk. In verschillende voorspellingsmethoden wordt de mate van verwantschap van de ouders gebruikt. Deze is onder andere bepaald met behulp van een relatief nieuw type moleculaire merkers: AFLPs. Europese tweerijige zomergerst (*Hordeum vulgare* L.) is gebruikt als modelgewas. Om betrouwbare en generaliseerbare resultaten te verkrijgen zijn 20 kruisingspopulaties, elk bestaand uit 48 inteeltlijnen (RILs: 'recombinant inbred lines'), samen met hun ouderlijnen beproefd in tienrijige veldjes in 7 milieus, verdeeld over twee jaar. De prestaties van de nakomelingen zijn waargenomen voor vier landbouwkundige eigenschappen: plantlengte, bloeitijdstip, duizendkorrelgewicht en korrel-opbrengst.

In hoofdstuk 2 wordt de genetische similariteit *gs*, gebaseerd op AFLP-merkers, vergeleken met andere verwantschapsmaten. De *gs* vertoont een zwakke correlatie met de 'coefficient of coancestry' f. Deze maat is gebaseerd op de afstamming van de ouderlijnen. Er is geen correlatie gevonden met de morfologische afstand *md*, die berekend is op basis van 25 morfologische eigenschappen. De nauwkeurigheid van de correlatieschattingen is bepaald door het nemen van 'bootstrap' steekproeven uit de populatie van ouderlijnen, waarbij door trekking met teruglegging een groot aantal 'bootstrap' datasets gecreëerd worden. Aan de hand van de daaruit berekende correlaties kan de onnauwkeurigheid van de correlatieschatting onderzocht worden. Voor wat betreft de toepassing van de verwantschapsmaat *gs*, blijkt dat deze geschikt is om enkele hoofd-ecotypes (bijv. wintergerst/ zomergerst, tweerijig/zesrijig) in gerst te onderscheiden. Ook blijken AFLPs bruikbaar voor rasidentificatie, zelfs al zijn ze gebaseerd op slechts één primercombinatie.

In hoofdstuk 4 wordt de bruikbaarheid van de verwantschapsmaten *gs*, *f* en *md* voor de voorspelling van de variantie van een nakomelingschap onderzocht. Correlaties tussen $1-f$ en de variantie van een RIL-populatie zijn meestal positief, maar niet significant. Slechts 10 oudercombinaties hadden betrouwbare afstammingsgegevens en konden voor deze diversiteitsmaat worden gebruikt. De correlaties tussen *md* en de variantie onder de nakomelingen zijn niet significant. De correlaties tussen $1-gs$ en de nakomelingschapsvariantie voor de verschillende eigenschappen zijn in het algemeen positief, maar zelden significant. Het vermoeden bestaat dat de AFLP-merkers geen goede representatie zijn van de genoomposities van de genen voor de verschillende eigenschappen. Dit zou de genoemde zwakke correlaties kunnen verklaren. Vervolgens is er nog een verwantschapsmaat onderzocht: de genetische afstand *agd* op basis van landbouwkundige eigenschappen in de verschillende milieus. De correlatie tussen *agd* en de variantie van de nakomelingen is meestal positief en soms significant, in het bijzonder wanneer de planteigenschappen voor de berekening van de *agd* tussen de ouders en de RIL-variantie hetzelfde zijn. Echter, deze correlaties berusten voornamelijk op verschillen tussen kruisingen in het aantal uitsplitsende hoofdgenen ('major genes') en hun effecten. Deze hoofdgenen zijn genen met grote effecten op een eigenschap en kunnen vaak 'op 't oog' beselecteerd worden. De polygenvariantie ('minor gene'variantie) die overblijft na een selectie op de gewenste allelen van uitsplitsende hoofdgenen, wordt niet betrouwbaar voorspeld door de *agd*. Tenslotte blijken gecombineerde verwantschapsmaten de hoogste correlaties te geven met de variantie van een nakomelingschap. Dit is zeker het geval wanneer voor verwante en onverwante ouderparen verschillende combinaties gebruikt worden. $1-gs$ en $1-f$ lijken meer betrouwbaar bij sterk verwante oudercombinaties, terwijl *agd* en *md* meer betrouwbaar lijken te zijn bij onverwante oudercombinaties. Echter, de correlaties van gecombineerde verwantschapsmaten met RIL-variantie zijn niet groot genoeg om betrouwbare variantievoorspellingen mogelijk te maken.

In hoofdstuk 5 worden twee varianten van de op AFLP-merkers gebaseerde genetische similariteit *gs* onderzocht. Deze varianten zijn mogelijk betere AFLP-representaties van de genoomposities van de genen voor de onderzochte eigenschappen. In de eerste variant wordt de oververtegenwoordiging van genoomposities als gevolg van geclusterde merkers geëlimineerd op basis van genetische kaartinformatie. Bij de berekening van de genetische afstand *mgd* worden de merkers gewogen voor de merkerdichtheid in de buurt van hun kaartpositie. De correlatie van de variantie van de nakomelingschap met deze *mgd* is echter net zo zwak als met $1-gs$, waarbij de merkers ongewogen zijn. De vooruitzichten voor een variantievoorspelling gebaseerd op *mgd*-schattingen zijn slecht. In de tweede variant wordt de genetische afstand *sgd* berekend op basis van merkers die geselecteerd zijn vanwege hun sterke associatie met een eigenschap in de ouderpopulatie. Het toegepaste selectiecriterium is de overschrijdingskans van de F-toets voor de onderzochte merker in een variantieanalyse van de oudergegevens.

Verschillende merkerselectieprocedures en selectiedrempels zijn onderzocht. De correlatie tussen *sgd* en RIL-variantie is het hoogst als er een reeks van ANOVA-modellen wordt gebruikt voor de merkerselectie. Deze modellen bevatten elk het effect van de onderzochte merker plus het effect van een van de andere beschikbare merkers. De gemiddelde overschrijdingskans voor de onderzochte merker in deze modellen wordt dan gebruikt als selectiecriterium. Voor verschillende eigenschappen is steeds een optimum selectiedrempel voor de gemiddelde overschrijdingskans gevonden van 0.005. Alleen bij bloeitijdstip bleek het nuttig om rekening te houden met merker×milieu-interactie. Echter, de correlaties tussen de *sgd* en de variantie van een nakomelingschap zijn niet hoog genoeg voor de verschillende eigenschappen om een betrouwbare variantievoorspelling te doen in de praktijk.

In hoofdstuk 3 wordt een beproeving van de kruisingen in een vroege inteeltgeneratie (F4) beschreven. Deze is uitgevoerd in kleine drierijige veldjes op twee locaties. De daaruit verkregen gegevens worden onderzocht op hun bruikbaarheid voor een voorspelling van gemiddelde en variantie van de later te verkrijgen RIL-populatie. De gemiddelde korrelopbrengst van een RIL-populatie kan niet worden voorspeld door de gemiddelde korrelopbrengst van een F4-populatie. Dit is een gevolg van concurrentie tussen de genotypen in de kleine veldjes en genotype×milieu-interactie. Voor de gemiddelde korrelopbrengst over milieus op basis van tienrijige veldjes, blijkt de 'midparent value', de gemiddelde prestatie van beide ouders, een redelijke voorspeller van de gemiddelde prestatie van de RIL-nakomelingschap (r=0.71). Echter, bij de meeste kruisingen levert een voorspelling op basis van 'midparent value' een overschatting op van de gemiddelde korrelopbrengst van de RILs. Dit is waarschijnlijk het gevolg van scheve uitsplitsing en epistasie. De variantie van een F4-nakomelingschap laat een redelijke correlatie zien met de variantie tussen de RILs (r=0.62). Echter, dit verband berust op verschillen tussen kruisingen in het aantal uitsplitsende hoofdgenen en hun effecten. De overblijvende polygenvariantie vertoont slechts een zwakke correlatie met de variantie tussen F4-lijnen (r=0.41). Het combineren van de F4-voorspelling van gemiddelde en variantie leidt tot een slechte voorspelling van de kans van een lijn in de nakomelingschap om beter te presteren dan een zekere drempelwaarde. Selectie op korrelopbrengst van F4-lijnen binnen een kruisingspopulatie blijkt ook niet effectief. Dit is het gevolg van de reeds genoemde competitie tussen genotypen in kleine veldjes, een grote standaardfout voor de individuele lijnschattingen en genotype×milieu-interactie.

In hoofdstuk 6 wordt de voorspelling van de prestaties van een nakomelingschap in de aanwezigheid van genotype×milieu-interactie onderzocht. Als eerste worden de 'midparent values' voor drie stabiliteitsmaten gebruikt om de gemiddelde stabiliteit over milieus van een nakomelingschap te voorspellen. Er wordt een redelijke correlatie gevonden tussen 'midparent value' en het gemiddelde van de RIL-populatie voor de coëfficiënt van regressie van de opbrengst van een genotype op de milieu-index (r_s=0.69). Deze milieu-index is de gemiddelde

opbrengst van alle ouders in een bepaald milieu. Een zwakkere, maar wel significante, correlatie wordt gevonden voor de tweede stabilititeitsmaat, de gemiddelde gekwadrateerde afwijking van de bovengenoemde regressie (r_s=0.39). Voor de derde maat, de stabiliteitsvariantie is de correlatie tussen 'midparent value' en het gemiddelde van de RIL-populatie zwak en niet significant (r_s=0.17). Dit is het gevolg van een verschil tussen ouders en nakomelingen in de samenstelling van deze maat. We concluderen dat van de drie stabiliteitsmaten alleen de gemiddelde regressiecoëfficiënt van de nakomelingschap voorspelbaar is in de praktijk. Dit is mede het gevolg van de slechte herhaalbaarheid van de gemiddelde gekwadrateerde afwijking van de regressie. Vervolgens is de genotype×milieu-interactie verder onderzocht middels een AMMI-analyse waarin de interactie wordt beschreven door een beperkt aantal multiplicatieve parameters. Hoewel er behoorlijke correlaties bestaan tussen 'midparent values' en nakomelingschapsgemiddelden voor deze parameters, zijn de vooruitzichten voor een voorspelling van de genotype×milieu-interactie van nakomelingen beperkt. Het interactiepatroon van de ouders kan echter ook gebruikt worden om hun onderlinge verwantschap vast te stellen. Als similariteitsmaat gebruiken we de correlatiecoëfficiënt r_{GXE}. Deze wordt berekend als de correlatie tussen de residu-vectoren van twee ouders, waarbij de residuen afkomstig zijn van een variantieanalyse van de opbrengstgegevens van de ouders in de verschillende milieus. Het gebruikte ANOVA-model bevat additieve genotype en milieu effecten. r_{GXE} blijkt negatief gecorreleerd met de variantie voor gemiddelde opbrengst over milieus in de nakomelingschap (r_s=−0.63). We concluderen dat een variantievoorspelling op basis van deze verwantschapsmaat misschien praktisch bruikbaar is, maar dat nog verder onderzoek nodig is.

Concluderend wordt gesteld dat het voorspellen van het nakomelingschapsgemiddelde met behulp van 'midparent values' bruikbaar is in de veredelingspraktijk. De voorspelling van de variantie in een RIL-populatie is moeilijker, zeker wanneer er geen hoofdgenen uitsplitsen. Daarom stellen wij voor om verwantschapsmaten tussen ouders te gebruiken als 'indicatoren' voor nakomelingschapsvariantie. Deze omschrijving benadrukt het ondersteunende karakter van verwantschapsmaten bij de keuze van oudercombinaties voor een kruisings- en selectieprogramma. Verschillende onderzochte verwantschapsmaten zijn bruikbaar om een voorzichtig onderscheid te maken tussen uniforme en variabele kruisingsnakomelingschappen. De mate van bruikbaarheid, ofschoon afhankelijk van gewas en eigenschap, kan tot op zekere hoogte worden bepaald aan de hand van de sterkte van de in dit proefschrift beschreven correlatie tussen de verwantschapsmaat en de nakomelingschapsvariantie.

Nawoord

Eindelijk is het dan af, het proefschrift, of, zoals het onder AIO's heet, 'het boekje'. Toen ik bijna zes jaar geleden met mijn AIO-project begon, was het al duidelijk dat het wel een jaartje of vijf zou gaan duren. Die planning is dus bijna gehaald, of ruim, het is maar hoe je het bekijkt. In elk geval sprak de inhoud van het project mij erg aan, omdat het een toepassing van moleculaire merkers in de plantenveredeling betrof en het werk behoorlijk praktijkgericht was. Deze merkertechnologie was toen nog erg nieuw en het was niet geheel duidelijk wat je nu wel en niet kon met merkers in de veredeling. Ik had het geluk dat de AFLP-merkertechnologie net op tijd binnen bereik kwam, waardoor het aantal te bepalen merkers drastisch omhooggeschroefd kon worden. Dit is m.i. dan ook geen 'bottleneck' geweest in het project. De 'bottleneck' was, zoals in veel veredelingsprojecten, de beschikbare capaciteit voor veldproeven. Desondanks is er nog niet eerder op een dergelijke omvangrijke schaal gekeken naar het verband tussen genetische afstand tussen ouderlijnen op basis van moleculaire merkers en variatie in een kruisingsnakomelingschap. Het was dan ook wat teleurstellend dat de resultaten geen duidelijk verband gaven te zien tussen deze twee parameters. Uit nadere analyse van de gegevens bleek dat dit niet zozeer te wijten was aan onnauwkeurigheid van de resultaten, maar dat het verband waarschijnlijk niet zo sterk was als vooraf vermoed werd. Enkele theoretische studies ondersteunden deze conclusie. Daarmee was een hele bult werk van een hele hoop mensen niet voor niets. Veredelaars weten nu dat zomaar wat merkergegevens op een hoop gooien weinig voorspellende waarde heeft en dat een voorselectie van merkers noodzakelijk is. Waarschijnlijk kunnen lopende en toekomstige QTL-studies daar nog een ondersteunende bijdrage aan leveren. Verder is uit het onderzoek ook duidelijk geworden dat het gebruik van gegevens over de landbouwkundige eigenschappen van de ouders minstens net zo belangrijk is om een voorspelling te doen over hun kruisingsnakomelingschap. Tenslotte kan ik nog toevoegen dat ikzelf veel geleerd heb van dit hele project. Het geeft een goed gevoel om een planning voor vier jaar veldje voor veldje realiteit te zien worden, zelfs als er lampendieven op je planten gaan staan of als hagel je proefveld vervroegd dorst. Het meeste plezier heb ik beleefd aan de samenwerking met iedereen die aan het project heeft bijgedragen. Alleen gezamenlijk kun je een dergelijke klus op een succesvolle manier klaren.

Ik zou allereerst de initiatiefnemers van het project willen bedanken: professor Jan Parlevliet en Ies Bos. Mede dankzij hen is dit onderwerp gekozen en kreeg ik de kans om het onderzoek te doen. De bijdrage van Lianke Breekland mag hier niet onvermeld blijven, want zij inventariseerde mogelijke onderzoeksdoelen in samenwerking met een aantal granenkwekers en viste dit onderwerp eruit. Daarna heeft ze ook gezorgd voor het uitgangsmateriaal en een globale onderzoeksopzet. Ik kwam dus min of meer in een gespreid bed. Daarna is ze steeds bij het project betrokken gebleven als lid van de begeleidingscommissie.

De kweekbedrijven Cebeco Zaden, Vanderhave en Zelder ben ik veel dank verschuldigd. Zij waren betrokken bij de start van het project en zorgden mede voor interessant uitgangsmateriaal, waaronder enkele splinternieuwe rassen. Daarnaast heb ik veel profijt gehad van de deelname van de gerstkwekers in de begeleidingscommissie van dit project. Door hun commentaar werd de koppeling van de onderzoeksresultaten met de veredelingspraktijk beter mogelijk. Tenslotte heb ik in 1996 bij bovengenoemde bedrijven uitgebreide veldproeven neer kunnen leggen. Dit is een onmisbare bijdrage geweest voor de resultaten in dit proefschrift.

De onvolprezen wetenschappelijke begeleiding van mijn project heeft zich behoorlijk uitgebreid in de loop van de zes jaar. Ies Bos was, zoals gezegd, vanaf het begin erbij. Zijn grote betrokkenheid, zijn uitgebreide literatuurkennis en zijn gevoel voor heldere formuleringen leverden een belangrijke bijdrage aan dit proefschrift, getuige zijn co-promotorschap en twee co-auteurschappen. Piet Stam kwam in 1994 als hoogleraar bij de vakgroep en verraste mij in ons eerste gesprek met de mededeling dat hij het promotorschap inmiddels had overgenomen van professor Parlevliet, terwijl ik op het punt stond om hem daarnaar te vragen vanwege zijn achtergrond in de kwantitatieve genetica en zijn ervaring met het gebruik van moleculaire merker gegevens. Uit dit proefschrift mag blijken dat ik van zijn kwalititeiten ruimschoots en dankbaar gebruik heb gemaakt. Daarnaast heeft Piet een goed oog voor structuur in het verhaal: iets wat je als AIO wel eens een enkele keer uit het oog dreigt te verliezen in het woud van resultaten. Tenslotte wil ik hier Johan Dourleijn bedanken, die nog weer iets later dan Piet op de vakgroep arriveerde, maar vanaf het begin van het project al in de begeleidingscommissie betrokken was. Zijn statistische kennis en zijn kritische blik vormden een goede aanvulling voor het begeleidingsteam.

Bij de uitvoering van alle proeven heb ik veel assistentie gehad. In de eerste plaats wil ik Jan Waninge bedanken, die mij al vrij snel in het project kwam helpen. De enorme hoeveelheid praktijkervaring die Jan meebracht, is bijzonder waardevol geweest voor het hele project. Ook in het contact met anderen in de kas en op het veld heb ik van Jan veel geleerd. Zijn vervroegde uittreding betekende dan ook een groot gemis voor het project en voor de vakgroep. Vervolgens wil ik Herman Masselink bedanken. Elk jaar kwam ik met grotere proeven en altijd weer wist hij het vele werk dat eraan vast zat goed te organiseren. Gelukkig kon hij daarbij rekenen op de ervaring van zijn medewerkers, eerst bij de vakgroep en daarna bij Unifarm. Ik wil dan ook iedereen die zijn steentje heeft bijgedragen aan de uitvoering van de proeven op het veld en in de kas van harte bedanken. Die dank geldt zeker ook de mensen in Flevoland, zowel bij de Minderhoudhoeve als bij de Broekemahoeve. De plezierige samenwerking maakte het doen van proeven op afstand tot een eenvoudig karweitje. Daarnaast wil ik de mensen op bovengenoemde kweekbedrijven bedanken voor hun bijdrage aan een serie prima proeven.

Naast de proeven in Nederland heb ik mijn materiaal ook bij het Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) in Mexico op het veld gehad. Ik wil professor

Parlevliet bedanken voor zijn bemiddeling om deze noodzakelijke vermeerderingsstap tot stand te brengen. Daarnaast bedank ik Hugo Vivar (Hugo, muchos gracias!), Leon Broers en het uitvoerders op het CIMMYT voor de prima begeleiding van de proef in Mexico. Daarnaast wil ik Leon en Ursula van harte bedanken voor hun gastvrijheid, ook namens Ingrid. We hebben het erg naar onze zin gehad bij jullie. Ook wil ik Maarten van Ginkel bedanken voor zijn enthousiaste rondleiding en de prettige samenwerking die volgde op mijn bezoek aan Obregon.

Ook op het lab is heb ik het nodige werk verricht voor dit project. I would like to thank Xiaoquan Qi for his help in teaching me how to apply the AFLP-technique in barley. Daarnaast wil ik Herman van Eck bedanken voor al zijn uitleg over het draaien van AFLPs. Ook al die andere mensen op het lab die mij hebben bijgestaan met adviezen over het verkrijgen van zuiver DNA uit gerst wil ik bedanken. Zonder al die hulp was het mij nooit gelukt om al die bandjes te verzamelen.

Tenslotte was er nog heel wat data-processing nodig om de gegevens op een rijtje te krijgen. Ik zou Theo van Hintum willen bedanken voor zijn database gerst-afstammingsgegevens en de software om die te analyseren. Ook zijn bijdrage in de begeleidingscommissie en zijn gezelschap tijdens het Barley Genetics congres in Canada waardeer ik bijzonder. Voor wat betreft de statistiek kon ik op een aantal specialisten terugvallen. Ik wil dan ook Scott Chasalow, Fred van Eeuwijk, Ritsert Jansen, en Rob Verdooren van harte bedanken voor hun meedenken.

Tijdens het project hebben verschillende studenten meegedraaid bij het onderzoek in het kader van hun afstudeervak. De resultaten van hun werk zijn niet allemaal concreet in dit proefschrift terechtgekomen, maar hebben wel bijgedragen aan de totstandkoming van de conclusies over kruisingsvoorspellingen. Ik wil dan ook Mark, Milika, Johannes en Derk bedanken voor hun werk en de prettige samenwerking.

Een beetje blikverbreding kan geen kwaad tijdens je AIO-schap. Een van de zaken die daarvoor zorgde was de onderzoekschool Productie-Ecologie. Ik heb met veel plezier meegedraaid in de methodologie-promovendigroep, die door Alfred Stein op enthousiaste wijze begeleid werd. Daarnaast heeft de Universiteitsraad mij twee jaar lang veel afleiding geboden. Ik wil dan het CvB, de griffie, het WAIOO en de raadsleden, in het bijzonder mijn fractiegenoten van de Centrale Lijst, van harte bedanken voor de goede samenwerking en de prettige sfeer.

Zoals inmiddels wel duidelijk is, doe je onderzoek niet alleen. Veel samenwerking is nodig om een goed resultaat tot stand te brengen. Ik ben blij dat dit altijd op een plezierige manier is verlopen. Hierbij is een goede werksfeer erg belangrijk. Ik wil dan ook graag alle collega's van de vakgroep hiervoor bedanken. De klaverjaspartijtjes, koffietafelgesprekken, Tour-de-France-pools, 'Gat'-bijeenkomsten, etc. verluchtigen het AIO-bestaan aanzienlijk. Mijn kamergenoten en Terpgenoten wil ik nog extra bedanken voor hun luisterende oor. Ik hoop dat ik dat ook voldoende gehoor aan hen heb kunnen bieden. Daarnaast wil ik de carpoolers uit

Breda en omstreken bedanken voor de gezelligheid onderweg. Met of zonder file, het was altijd weer prettig om samen het ritje heen en terug te maken.

Mijn ouders wil ik bedanken voor de ruimte en steun die ze mij hebben geboden om te gaan studeren en hun belangstelling voor mijn onderzoek. Door hen ben ik enthousiast geworden om iets op agrarisch gebied te gaan leren en daar heb ik zeker geen spijt van. Van mijn zussen heb ik geleerd dat er ook nog iets anders bestaat dan gerst, veredeling en landbouw. Als ik dat weer eens dreig te vergeten, word ik er door hen gelukkig altijd weer aan herinnerd.

Tenslotte wil ik jou bedanken, Ingrid. We hebben met elkaar de afgelopen zes jaar ondanks dit proefschrift er iets moois van kunnen maken. Of we nu bij elkaar woonden of op afstand, ik heb altijd je betrokkenheid en begrip gevoeld. Door jouw nuchtere opmerkingen was het voor mij een stuk makkelijker het werk te relativeren. Onze zondagse wandelingen en fietstochten waren een prima manier om het werk te vergeten en over belangrijkere dingen in het leven te praten. Ik hoop dat we dat zo in de toekomst blijven doen. Met zijn tweeën sta je tenslotte veel sterker dan alleen. Daarom een dikke papieren kus en knuffel van mij.

Curriculum vitae

Johannes (Johan) Wilhelmus Schut werd geboren op 21 April 1968 te Biddinghuizen, dat toentertijd deel uitmaakte van het openbaar lichaam Zuidelijke IJsselmeerpolders. In 1971 verhuisde hij met zijn ouders naar het rozendorp Lottum. Hij behaalde in 1986 het Atheneum-B-diploma aan het Collegium Marianum te Venlo en begon vervolgens met zijn studie Plantenveredeling aan de Landbouwhogeschool te Wageningen. In 1988 onderbrak hij zijn studie om de functie van Ab-actis/Quaestor te bekleden in het bestuur 88-II van de KSV Sint Franciscus Xaverius. Nadat hij zijn studie weer opgepakt had, liep hij in 1990 stage in het brouwgerstveredelingsprogramma van Prof. B.L. Harvey aan de University of Saskatchewan te Saskatoon, Canada. Verder deed hij afstudeeronderzoek bij de vakgroepen Wiskunde (Wiskundige Statistiek), Erfelijkheidsleer, Plantenveredeling (Selectiemethoden) en Theoretische Productie-Ecologie. In augustus 1992 studeerde hij cum laude af, waarna hij in november begon als Assistent in Opleiding (AIO) in het project 'The predictive value of the degree of relationship of parent lines on the results of selection'. Het onderzoeksproject werd uitgevoerd bij de vakgroep Plantenveredeling. De meeste resultaten van dit onderzoek zijn beschreven in dit proefschrift. Tijdens zijn AIO-schap maakte hij van 1993 tot 1995 deel uit van de Universiteitsraad van de Landbouwuniversiteit, waarbij hij zich bezighield met Onderwijs en Onderzoek, Financiën en Planning, Studentenzaken en enkele specifieke AIO-onderwerpen. Sinds augustus 1998 werkt hij als slaveredelaar bij Rijk Zwaan te Hendrik-Ido-Ambacht.