Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet Angewandte Genetik und Pflanzenzüchtung
Prof. Dr. A.E. Melchinger

Genetic Diversity, Population Structure, and Linkage Disequilibrium in the Context of Genome-wide Association Mapping of Northern Corn Leaf Blight Resistance

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften
vorgelegt
der Fakultät Agrarwissenschaften

von
Ingénieur Agronome (FR)
Delphine Van Inghelandt
aus Seclin, Frankreich

Passau
2012

Die vorliegende Arbeit wurde am 19.10.2011 von der Fakultät Agrarwissenschaften der Universität Hohenheim als "Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)" angenommen.

Tag der mündlichen Prüfung: 16.02.2012

1. Prodekan:	Prof. Dr. A. Fangmeier
Berichterstatter, 1. Prüfer:	Prof. Dr. A.E. Melchinger
Mitberichterstatter, 2. Prüfer:	Prof. Dr. H.-P. Piepho
3. Prüfer:	Prof. Dr. A. Charcosset

Contents

---

[1] Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010) Theor Appl Genet 120:1289–1299

[2] Van Inghelandt D, Reif JC, Dhillon BS, Flament P, Melchinger AE (2011) Theor Appl Genet 123:11–20

[3] Van Inghelandt D, Melchinger AE, Martinant JP, Stich B (2012) BMC Plant Biol 12:56

Abbreviations

AFLP	amplified fragment length polymorphism
AM	association mapping
AMMSP	association mapping in multiple segregating populations
CV	coefficient of variation
D	gene diversity
f	coancestry coefficient
F_{ST}	population fixation index
FT	flowering time
GEBV	genomic estimated breeding value
GS	genomic selection
JLAM	joint linkage and association mapping
LD	linkage disequilibrium
LM	linkage mapping
MAS	marker assisted selection
MRD	modified Roger's distance
NAM	nested association mapping
NCLB	northern corn leaf blight
NCLB_{FT}	NCLB corrected for flowering time
N_e	effective population size
NLR	non-linear regression
QTL	quantitative trait loci
R	squared correlation coefficient, measure of LD
SNP	single nucleotide polymorphism
SSR	simple sequence repeat
SSS	Stiff Stalk Synthetic

1. General Introduction

Maize (*Zea mays* L.) is a very important crop for feed, energy, and food production throughout the world. It is a monoecious species with separated male (tassel) and female (ear) organs. This makes the genetic diversity of maize remarkable, because it outcrosses easily. This diversity has been captured in various forms, including inbred lines, native landraces, and open-pollinated populations. Maize is therefore an excellent model for basic research (Candela and Hake 2008). Furthermore, maize was one of the pioneer species in the applied plant breeding area during the past 100 years, especially for the development of hybrid varieties. This inbred-hybrid concept is still considered as one of the greatest achievements in crop breeding (Hallauer and Miranda 1988) and is the variety type of choice for many species.

Population structure and genetic diversity

In hybrid breeding of maize, knowledge about genetic relationships among inbreds is useful for germplasm organization and cultivar protection (Melchinger et al. 1991; Bernardo 2002). In the context of germplasm organization, inbreds can be grouped according to their estimates of genetic similarity and assigned to heterotic pools. For plant variety protection, information on genetic distances among inbreds is important for the identification of essentially derived variety as well as for legal protection of germplasm

(Smith et al. 1995). Therefore, information about the genetic diversity and population structure in elite breeding material is of fundamental importance for the improvement of crops (Hallauer and Miranda 1988).

Various avenues have been suggested in the literature to achieve this goal. A widely used measure in this context was the coancestry coefficient f calculated from pedigree records, which is defined as the probability that two homologous genes drawn at random from two individuals are identical by descent (Malécot 1948). Nevertheless, pedigree records tracing back more than two generations are rare. A further shortcoming is that some founder inbreds of heterotic pools were derived from open-pollinated populations. Hence, calculation of f is often not feasible or dubious in maize (Lübberstedt et al. 2000).

With the appearance of molecular markers in the late 1980s, new alternatives became available in agriculturally important crops to assess genetic characteristics on a level which was not possible before (Melchinger and Gumber 1998). Such DNA markers can be applied to dissect the relationship between genotypes and help to classify them, detect the impact of selection in these crops, and quantify the level of genetic diversity in their genome. Furthermore, molecular markers can be used to dissect quantitative traits into their underlying genetic factors called quantitative trait loci (QTL).

Extent and genome-wide distribution of linkage disequilibrium

Beside linkage mapping (LM) as a routine tool for the identification of QTL in plants, association mapping (AM) became a powerful complement for understanding the genetic basis of complex traits (Yu et al. 2008). These two methods exploit the linkage disequilibrium (LD) between genes coding for a trait and closely linked markers. LD, also known as gametic phase

disequilibrium, is the non-random association between alleles at different loci.

In LM studies, the extent of LD can be calculated theoretically because these are carried out with populations constructed from bi-parental crosses. In contrast, AM makes use of the LD resulting from many generations of historical recombinations in germplasm with unknown relatedness. In this case, the extent of LD is affected by many genetic factors such as mutation, recombination, selection, migration, and mating pattern (Flint-Garcia et al. 2003) and can only be determined empirically. Therefore, the germplasm of interest must be examined with respect to the extent of LD and its genomic distribution to determine the prospects of genome-wide association mapping (Rafalski 2002; Kim et al. 2007).

The relationship of LD with physical or genetic distance is highly variable across species, marker systems, and genome regions (for review see Flint-Garcia et al. 2003; Rafalski and Morgante 2004; Yu and Buckler 2006). In maize, different germplasm sets such as indigenous landraces, exotic materials, and lines from public breeding programs have been examined for LD in certain genomic regions. Yan et al. (2009) used 632 inbred lines from temperate, tropical, and subtropical public breeding programs for a whole genome LD scan. In this study, LD decay distances differed among chromosomes and ranged from 1 to 10 kb, and it was much higher in temperate than tropical and subtropical germplasm. Nevertheless, knowledge of LD over large map distances is still limited particularly in commercial maize germplasm (Veyrieras et al. 2007).

Comparison of simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers for their use in plant breeding

Until now, SSR markers have been the most widely used DNA marker type to characterize germplasm collections of crops because of their easy use, relatively low price, and high degree of polymorphism provided by the large number of alleles per locus (Vignal et al. 2002). However, for a whole genome AM approach, a high marker density is required in germplasm with a rapid LD decay. SSR markers may lack the density needed for AM studies (Ching et al. 2002).

More recently, SNP markers received high attention because they occur at much higher frequency in the genome than SSRs, and, therefore could provide the required high marker density. Furthermore, their genotyping can be easily automated. However, most SNPs are bi-allelic, and, thus, have a lower information content. Given the advantages and disadvantages of both marker systems, their usefulness in different fields of application must be compared.

When assessing the repeatability of genotyping results and proportion of missing data for SSR and SNP markers, Jones et al. (2007) found a clear advantage for SNPs. In contrast, Hamblin et al. (2007) investigated the usefulness of 89 SSRs versus 847 SNPs for assessing relatedness and evaluating genetic diversity in a set of public maize inbreds and found that SSRs performed better with respect to the assignment of inbreds to sub-populations. These authors suggested that based on their study a considerable higher number of SNP markers might be required in order to have an equivalent discriminating power as with SSRs. Nevertheless, to our knowledge, no earlier study examined this issue, especially in elite maize germplasm, nor considered the differences in costs for genotyping SSRs and SNPs.

On the other hand, Stich et al. (2006) compared multi-allelic (SSR) and bi-allelic (amplified fragment length polymorphism; AFLP) markers to investigate LD in a set of 72 central European maize inbred lines and concluded that SSRs should be preferred over AFLPs except in populations having a long history of recombination. In a set of 102 maize inbreds representing a broad cross-section of public breeding germplasm from temperate and tropical regions, Remington et al. (2001) compared the extent of LD in six genes based on SNPs versus genome-wide distributed SSRs and observed that SSRs revealed stronger evidence of LD than SNPs. Nevertheless, a direct comparison of the extent of LD between genome-wide distributed SSRs and SNPs in commercial maize breeding germplasm is lacking.

Genome-wide association mapping of northern corn leaf blight (*Setosphaeria turcica*) resistance

Setosphaeria turcica (anamorph *Exserohilum turcicum*, formerly known as *Helminthosporium turcicum*) is a fungal pathogen that causes northern corn leaf blight (NCLB) in maize. Infections of maize with NCLB before silking can cause grain yield losses of more than 50%, which are accompanied by a reduction in feed value and the predisposition of infected plants to stalk rot (Fajemisin and Hooker 1974).

Plants have evolved qualitative and quantitative resistance to combat pathogens. For NCLB, qualitative resistances have been identified and called *Ht* genes (for *Helminthosporium turcicum*). These monogenic resistances have been backcrossed into a number of widely used inbred lines, where they showed partial dominance and expression dependent on the genetic background (Welz 1998). Furthermore, the expression of the *Ht* genes is modified by the environment, particularly temperature and light intensity (Thakur et

al. 1989a). In addition, qualitative resistances conferred by single genes such as the *Ht* genes tend to be overcome by new, virulent races of *Setosphaeria turcica* (*e.g.* Thakur et al. 1989b; Pataky et al. 1991; Windes and Pedersen 1991). All these aspects limit the practical value of the *Ht* genes and have hampered their use in maize breeding programs.

Quantitative resistances are considered to be oligo- or polygenically inherited and, thus, partially as well as moderately effective, but race unspecific and durable (Poland et al. 2009). Due to the latter two properties, quantitative resistances are today considered more useful in a breeding context than qualitative resistances. In agreement with this notion, the majority of disease resistances deployed in elite varieties of maize are quantitative (Wisser et al. 2006). However, identification of genes conferring quantitative resistance is much more challenging than identifying resistance genes, owing to their smaller phenotypic effects.

Various studies have been conducted to map QTL for resistance to NCLB (for review see Wisser et al. 2006). All of them were linkage mapping studies using different types of progenies such as F2 or F3 generations, BC1 generations, or populations of near-isogenic lines or recombinant inbred lines. Owing to the large confidence intervals of QTL and a restricted allelic sampling in the two parental genotypes, however, the results of linkage mapping studies had so far little impact on resistance breeding (Wisser et al. 2006). Very recently, NCLB resistance in maize was dissected using the nested association mapping (NAM) population (Poland et al. 2011), which offers the advantage of a higher mapping resolution and a broader allelic sampling than the above mentioned linkage mapping studies. Nevertheless, population-based association mapping has the potential of resulting in an even higher mapping resolution and broader allelic sampling compared to NAM (Ersoz et al. 2009). To our knowledge, however, no genome-wide population-based association mapping study has been yet conducted for NCLB resistance in maize.

Resistance genes identified by linkage or association mapping might affect the disease either directly or indirectly. Genes affecting plant growth and development or time to flowering fall in the latter class (Wisser et al. 2006). Especially for diseases caused by necrotrophic pathogens such as *Setosphaeria turcica*, which are more severe on senescing leaf tissue after anthesis, a relationship between plant disease resistance and flowering time (FT) might be expected (Wisser et al. 2006). In contrast to these linkage mapping studies, our association analysis will allow to discriminate with a high mapping resolution between pleiotropy and linkage of QTL for NCLB resistance and FT (cf. Stich et al. 2008).

Objectives

The goal of this thesis research was to set up the stage and perform association mapping in elite maize breeding populations for NCLB resistance. In particular, the objectives were to

1. examine the population structure and the genetic diversity in elite maize breeding lines based on SSR markers;
2. compare these results with those obtained from SNP markers;
3. examine the extent of LD with SSRs and SNPs in 1 537 commercial maize inbred lines belonging to four heterotic pools;
4. compare the LD patterns determined by using these two marker types,
5. evaluate the number of SNP markers needed to perform genome-wide association analyses;
6. identify chromosomal regions affecting FT and NCLB resistance using genome-wide association mapping;

7. examine the epistatic interactions of the identified chromosomal regions with the genetic background on an individual marker basis; and

8. dissect the correlation between NCLB resistance and FT.

References

Bernardo R (2002) Breeding for quantitative traits in plants. Stemma Press, Woodbury, p 41, p 249

Candela H, Hake S (2008) The art and design of genetic screens: Maize. Nature Rev Genet 9:192–203

Ching A, Caldwell KS, Jung M, Dolan M, Smith OSH, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:1–14

Ersoz ES, Yu J, Buckler ES (2009) Applications of linkage disequilibrium and association mapping in maize. Springer Verlag, Berlin, pp 173–195

Fajemisin JM, Hooker AL (1974) Predisposition to *Diplodia-zeae* stalk rot in corn affected by 3 *Helminthosporium* leaf blights. Phytopathology 64:1496–1499

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Hallauer AR, Miranda JBF (1988) Quantitative Genetics in Maize Breeding 2nd edn. Iowa State University Press, Ames

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE 2(12):e1367

Jones ES, Sullivan H, Bhattramakki D, Smith JSC (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (Zea mays L.). Theor Appl Genet 115:361–371

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nature genetics 39:1151–1155

Lübberstedt T, Melchinger AE, Dule C, Vuylsteke M, Kuiper M (2000) Relationships among early European maize inbreds: IV Genetic diversity revealed with AFLP markers and comparison with RFLP, RAPD, and pedigree Data. Crop Sci 40:783–791

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

Melchinger AE, Messmer MM, Lee M, Woodmana WL, Lamkey KR (1991) Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci. 31:669–678

Melchinger AE, Gumber RK (1998) Overview of heterosis and heterotic groups in agronomic crops. In: Lamkey KR and Staub JE (ed.) Concepts and breeding of heterosis in crop plants. CSSA Spec. Publ. 25. CSSA, Madison, pp 29–44

Pataky JK, Carson ML, Mosely PR (1991) Race 23N of *Exserohilum turcicum* in Florida. Plant Disease 75:813

Poland JA, Balint-Kurti PJ, Wisser RJ, Pratt RC, Nelson RJ (2009) Shades of gray: the world of quantitative disease resistance. Trends Plant Sci 14:21–29

Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proc Natl Acad Sci USA 108:6893–6898

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Rafalski A, Morgante M (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20:103–111

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci 98:11479–11484

Smith JSC, Ertl DS, Orman BA (1995) Identification of maize varieties. In: Wrigley CW (ed.) Identification of food grain varieties. Am Assoc Cereal Chemists, St Paul, pp 253–264

Stich B, Maurer HP, Melchinger AE, Frisch M, Heckenberger M, van der Voort JR, Peleman J, Sørensen AP, Reif JC (2006) Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Mol Breed 17:217–226

Stich B, Melchinger AE, Heckenberger M, Möhring J, Schechert A, Piepho HP (2008) Association mapping in multiple segregating populations of sugar beet (*Beta vulgaris* L.). Theor Appl Genet 117:1167–1179

Thakur RP, Leonard KJ, Leath S (1989a) Effects of temperature and light on virulence of *Exserohilum turcicum* on corn. Phytopathology 79:631–635

Thakur RP, Leonard KJ, Jones RK (1989b) Characterization of a new race of *Exserohilum turcicum* virulent on corn with resistance gene HtN. Plant Dis 73:151–155

Veyrieras JB, Camus-Kulandaivelu L, Gouesnard B, Manicacci D, Charcosset A (2007) Bridging genomics and genetic diversity: Linkage disequilibrium structure and association mapping in maize and other cereals. Crop Sci 47:60–71

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305

Welz HG (1998) Genetics and epidemiology of the pathosystem *Zea mays/Setosphaeria turcica*. Habilitation thesis. University of Hohenheim, Stuttgart

Windes JM, Pedersen WL (1991) An isolate of *Exserohilum turcicum* virulent on maize inbreds with resistance gene HtN. Plant Disease 75:430

Wisser RJ, Balint-Kurti PJ, Nelson RJ (2006) The genetic architecture of disease resistance in maize: A synthesis of published studies. Phytopathology 96:120–129

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE 4:e8451

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers

Delphine Van Inghelandt, Albrecht E. Melchinger, Claude Lebreton and Benjamin Stich

D. Van Inghelandt, A. E. Melchinger: Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

D. Van Inghelandt, C. Lebreton: Limagrain Verneuil Holding, Ferme de l'Étang, BP3, 77390 Verneuil l'Étang, France

B. Stich: Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

Theoretical and Applied Genetics (2010) 120:1289-1299

The original publication is available at www.springerlink.com

Abstract

Information about the genetic diversity and population structure in breeding material is of fundamental importance for the improvement of crops. The objectives of our study were to (i) examine the population structure and the genetic diversity in elite maize germplasm based on simple sequence repeat (SSR) markers, (ii) compare these results with those obtained from single nucleotide polymorphism (SNP) markers, and (iii) compare the coancestry coefficient calculated from pedigree records with genetic distance estimates calculated from both marker types. Our study was based on 1,537 elite maize inbred lines genotyped with 359 SSR and 8,244 SNP markers. The average number of alleles per locus, of group specific alleles, and the gene diversity (*D*) were higher for SSRs than for SNPs. Modified Roger's distance (MRD) estimates and membership probabilities of the STRUCTURE matrices were higher for SSR than for SNP markers but the

germplasm organization in four heterotic pools was consistent with STRUCTURE results based on SSRs and SNPs. MRD estimates calculated for the two marker systems were highly correlated (0.87). Our results suggested that the same conclusions regarding the structure and the diversity of heterotic pools could be drawn from both types of markers. Furthermore, although our results suggested that the ratio of the number of SSRs and SNPs required to obtain MRD or *D* estimates with similar precision is not constant across the various precision levels, we propose that between 7 and 11 times more SNPs than SSRs should be used for analyzing population structure and genetic diversity.

Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm

Delphine Van Inghelandt, Jochen C. Reif, Baldev S. Dhillon, Pascal Flament and Albrecht E. Melchinger

D. Van Inghelandt, B. S. Dhillon, A. E. Melchinger: Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

D. Van Inghelandt, P. Flament: Limagrain Europe, Site ULICE, BP 173, 63204 Riom Cedex, France

J. C. Reif: State Plant Breeding Institute, University of Hohenheim, 70593 Stuttgart, Germany

Theoretical and Applied Genetics (2011) 123:11-20

The original publication is available at www.springerlink.com

Abstract

Association mapping is based on the linkage disequilibrium (LD) resulting from historical recombinations and helps understanding the genetic basis of complex traits. Many factors affect LD and, therefore, it must be determined empirically in the germplasm under investigation to examine the prospects of successful genome-wide association mapping. The objectives of our study were to (i) examine the extent of LD with simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers in 1,537 commercial maize inbred lines classified in four heterotic pools, (ii) compare the LD patterns determined by these two marker types, (iii) evaluate the number of SNP markers needed to perform genome-wide association analyses, and (iv) investigate temporal trends of LD. Mean values of the squared correlation coefficient ($\overline{R}$) were almost identical for unlinked, linked, and adjacent SSR marker pairs. In contrast, $\overline{R}$ values were the lowest for the unlinked SNP loci and the highest for the SNPs within amplicons. LD decay varied across the different heterotic pools and the individual chromosomes. The SSR markers

employed in the present study are not adequate for association analysis, because of insufficient marker density for the germplasm evaluated. Based on the LD decay in the various heterotic pools, we would need between 4,000 and 65,000 SNP markers to detect with a reasonable power associations with rather large quantitative trait loci (QTL). A much higher marker density is required to identify QTL with smaller effects. However, not only the total number of markers but also their distributions among and along the chromosomes are primordial for undertaking powerful association analyses.

Genome-wide association mapping of flowering time and northern corn leaf blight (Setosphaeria turcica) resistance in a vast commercial maize germplasm set

Delphine Van Inghelandt, Albrecht E. Melchinger, Jean-Pierre Martinant and Benjamin Stich

D. Van Inghelandt, A.E. Melchinger: Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

J.P Martinant: Limagrain Europe, Site Ulice, BP 173, 64204 Riom Cedex, France

B. Stich: Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany

D. Van Inghelandt, Current address: Limagrain GmbH, Breeding Station, Schoenburg 6, 94060 Pocking, Germany

BMC Plant Biology (2012) 12:56

The original publication is available at www.biomedcentral.com

Abstract

Setosphaeria turcica is a fungal pathogen that causes in maize a serious foliar disease named northern corn leaf blight (NCLB). In order to unravel the genetic architecture of the resistance against this disease, a vast association mapping panel comprising 1,487 European maize inbred lines was used to (i) identify chromosomal regions affecting flowering time (FT) and northern corn leaf blight (NCLB) resistance, (ii) examine the epistatic interactions of the identified chromosomal regions with the genetic background on an individual molecular marker basis, and (iii) dissect the correlation between NCLB resistance and FT. The single marker analyses performed for 8,244 single nucleotide polymorphism (SNP) markers revealed seven, four, and four SNP markers significantly ($\alpha = 0.05$, amplicon wise Bonferroni correction) associated with FT, NCLB, and NCLB resistance corrected for FT, respectively. These markers explained individually

between 0.36 and 14.29% of the genetic variance of the corresponding trait. The very well interpretable pattern of SNP associations observed for FT suggested that data from elite plant breeding programs can be used to dissect polygenic traits. This in turn indicates that the associations identified for NCLB resistance might be successfully used in marker-assisted selection programs. Furthermore, the associated genes are also of interest for further research concerning the mechanism of resistance to NCLB and plant diseases in general, because in the literature some of the associated genes have not been mentioned in this context so far.

5. General Discussion

Association mapping (AM) is a promising method to genetically dissect complex traits (Rafalski 2010). It refers to the analysis of statistical association between genotypes (molecular markers) and phenotypes (traits). It is therefore closely related to linkage mapping (LM). However, in contrast to LM, which uses on purpose created segregating populations, AM utilizes the natural genetic diversity and linkage disequilibrium (LD) present in a diverse germplasm set.

Population structure and genetic diversity

An important consequence of using a diverse panel of germplasm for AM is the possible presence of population structure or familial relatedness due to the obscure or complex population history (Yu et al. 2006). Population structure and familial relatedness, if unaccounted for, can lead to an increased rate of spurious marker-trait associations, because they create LD between unlinked loci (Wright and Gaut 2005). Therefore, these factors have constrained the use of AM in plant genetics (Flint-Garcia et al. 2003). These aspects are even more important in plant breeding populations, like examined in the frame of this thesis, that have experienced a high selection pressure along the breeding process. As a consequence, it was of high importance to study the population structure and genetic diversity in the germplasm under consideration before being able to perform AM.

Molecular markers are widely used in breeding programs from private companies for variety protection and to perform population structure and genetic distance analyses. Several marker types are available for these purposes. One important concern in AM is whether adequate markers are used to account for this genetic relatedness (Yu et al. 2009). Therefore, a comparison of the two maker types most frequently used by maize breeders, namely SSRs and SNPs, with respect to their ability to perform population structure and genetic distances analyses was performed.

Modified Roger's distance (MRD) estimates calculated for the two marker systems were highly correlated and the trends observed for the gene diversity (D) of the different heterotic pools were identical. The assignment to a STRUCTURE sub-group based on SSRs and SNPs was for 97% of the inbreds identical, when using the highest membership probability criterion. Furthermore, the sub-groups identified with this procedure were in accordance with the heterotic pools as well as with the clusters revealed by PCoA. These results suggested that the same conclusions regarding the structure and the diversity of heterotic pools can be drawn from both markers types.

However, the correlation between the MRD estimates calculated from SSR and SNP data varied slightly between the different heterotic pools. Within the Lancaster and Flint pools, for which a high D was observed, a slightly lower correlation was observed than for the Stiff Stalk (SSS) and Iodent pools. An even lower correlation was found for the MRD between heterotic pools calculated from SSRs and SNPs. These two findings are in accordance with the results of Hamblin et al. (2007) and Jones et al. (2007) and can be explained by the fact that genetic distances between related genotypes are more precisely estimated than those between unrelated genotypes.

In addition to the trends regarding genetic distance and diversity estimates, the variance associated with them is an important criterion for marker applications in plant breeding. Therefore, we compared the coefficient of variation (CV) of MRD estimates calculated from SSRs and SNPs using a

bootstrap procedure. Considering the similar genotyping costs for both data sets, our results suggested that based on the same budget for genotyping, MRD can be more precisely estimated with SNPs than with SSRs. Under the assumption that a CV of 1% is sufficient for the estimation of genetic distances, our results suggested that about 270 SSRs or 3 150 SNPs (ratio SSR:SNP 1:11) are required to reach this precision.

Despite the general trends of the results for the different heterotic pools were identical for the SSR and SNP markers, differences were observed for the absolute values of the average number of alleles per locus, group specific alleles, and D, which were higher for SSRs than for SNPs. Accordingly, MRD estimates and membership probabilities of the STRUCTURE matrices were higher for SSR than for SNP markers. These findings are due to the fact that the SNPs are usually bi-allelic (Vignal et al. 2002). Theoretical considerations suggests that the maximum gene diversity D observable with bi-allelic markers is 0.5, whereas for multi-allelic markers such as SSRs the maximum is 1. As a consequence of the homozygous germplasm analyzed, the expected total heterozygosity h_T is in our case identical to D. Therefore, the differences observed for the overall population fixation index F_{ST} calculated from SSRs and SNPs can be explained by the definition of F_{ST} (Wright 1965):

$$F_{ST} = \frac{(h_T - h_S)}{h_T}, \tag{1}$$

where h_S i the expected within-population heterozygosity. h_T is high due to the multi-allelism of the SSRs, and, thus, F_{ST} of SSRs is expected to be smaller than the one calculated from SNPs. One possibility to circumvent the problem of the different number of alleles between SSRs and SNPs would be to combine the single SNP markers to SNP haplotype blocks.

Combining SNP markers to SNP haplotypes

SNP haplotype blocks are a set of closely linked SNP single markers which are not easily separable by recombination and therefore tend to be inherited

together (Balding 2006). In human genetics, haplotype blocks were found to be more powerful discriminators than single SNPs between cases and controls in disease association studies. Another advantage is that evolutionary studies can be conducted with haplotypes. Furthermore, the use of SNP haplotypes instead of single SNP markers in association studies reduces the number of statistical tests to be carried out. Nevertheless, to be able to use SNP haplotype blocks in genetic analyses, the way of combining the SNPs has first to be defined.

For heterozygous individuals, the phase of the alleles is unknown, *i.e.* it is unknown whether the haplotype block was inherited from the father or mother, and, therefore, this has to be experimentally or statistically inferred. When based on genotyping the gametes of the parents or related members of the population, haplotype block building is very expensive (Balding 2006). To overcome this problem, statistical methods have been developed for haplotype construction and are implemented in different software (*e.g.* PHASE, SNPHAP, FASTPHASE) that perform quite well with high density SNPs and few missing data.

For the special case of maize breeding germplasm, usually homozygous individuals are used and genotyped and, thus, inferring the phase is not required. Nevertheless, even in this simple scenario, it is not straight forward to decide how many and which markers have to be combined. Jansen et al. (2003) specified a fixed block length of adjacent mapped markers. On the other side, Schrag et al. (2007) developed strategies to retain haplotype blocks showing strong LD inside the blocks. This procedure requires knowledge of the extent and distribution of LD in the germplasm under consideration.

Extent of bi- and three-dimensional LD

AM utilizes the historical recombination events at the population level and is built on the basis of the LD concept (Ersoz et al. 2007). Because AM

uses ancestral LD in a population, there is more time and opportunity for recombination to take place than in linkage based approaches using pedigree populations. This leads to a reduction in the extent of LD and, therefore, enables a higher mapping resolution in AM than in LM (Thornsberry et al. 2001). However, this is particularly the case when using populations of unrelated individuals, where sampled chromosomes are much more distant than in pedigree populations. In breeding germplasm as studied in the frame of this thesis, the individuals of a heterotic pool cannot be considered as completely unrelated, because the best genotypes has been recombined together along the breeding history. Therefore, it was of fundamental importance for AM with breeding germplasm to assess empirically the extent of LD in the population under consideration, and to find out the type and number of markers required to perform successfully AM.

Mean values of the squared correlation coefficient (R), a widely applied measure for bi-dimensional LD, were almost identical for unlinked, linked, and adjacent SSR marker pairs. This clearly indicated that the SSR marker density employed in this study was not adequate for association analysis. In contrast, R values were lowest for the unlinked SNP loci and highest for the SNPs derived from the same amplicons. This suggested a relationship between the extend of LD and genetic distance, which enable the use of this type of marker for AM.

Quantifying the LD decay with increasing genetic map distances in our germplasm will facilitate the design of a SNP chip for AM studies. In order to do so, the following non linear regression (NLR), which gives the expectation of R between adjacent sites using the model of Hill and Weir (1988), was fitted to our data:

$$E(R) = \frac{(10+C)}{(2+C)\times(11+C)} \times (1 + \frac{(3+C)\times(12+12C+C^2)}{n\times(2+C)\times(11+C)}), \qquad (2)$$

where n is the sample size and C the population recombination parameter (with $C{=}4N_e c$; N_e being the effective population size and c the recombina-

tion fraction between the loci pairs considered). *Ne* is the number of individuals in an idealized population that could show the same amount of dispersion of allele frequencies under random genetic drift (or the same amount of inbreeding) as the population under consideration (Wright 1931; Wright 1938). *Ne* empirically calculated with NLR on equation (2) for the SSS pool was 72, whereas this pool had been developed originally using 16 inbreds (Hagdorn et al. 2003). The estimated *Ne* may indicate a slight overestimation compared to the real effective population size, which is calculated as the harmonic mean of the population size over the generations. This slight overestimation can be due to violation of the assumption of drift-recombination equilibrium and a low level of mutations made by the model (Remington et al. 2001), as well as the action of selection during the breeding process and the introduction of new genetic material. Nevertheless, the estimates of *Ne* for the different heterotic pools represented the expected ratio between them (for instance *Ne*= 1723 for the Flint pool and *Ne*= 72 for the SSS pool), because we found that the Flint pool was the most and the SSS pool the least diverse. Furthermore, the curves obtained by plotting recombination fraction vs. the expected values of *R* matched closely the curves obtained from locally weighted regression smoothing of recombination fraction vs. the expected values of *R*. Consequently, the estimates of *Ne* obtained may not be precise, but have nonetheless utility for characterizing the rate of LD decay per chromosome in each heterotic pool and are therefore relevant for practical maize breeders.

LD decay estimated with the NLR varied strongly across the different heterotic pools and the individual chromosomes. Based on these different values of LD decay, we would need between 4 000 (for the SSS pool), 6 800 (for the Iodent pool), 17 000 (for the Lancaster pool), and 65 000 (for the Flint pool) SNP markers to detect with a reasonable power associations with QTL explaining at least 10% of the phenotypic variation in a population of 350-400 individuals (Yu et al. 2006) like in the present study. A much higher marker density is required to identify QTL with smaller effects.

I described above the pattern of bi-dimensional LD, *i.e.*, the LD between

pairs of markers along the chromosome, which is the most frequent pattern of LD studied. However, for the building of haplotype blocks from SNP markers, which has been discussed in the previous paragraph, considering the joint effect of a group of markers beyond the pairwise connections could be an advantage compared to bi-dimensional LD (Nielsen et al. 2004). An illustrative example of how this three-dimensional LD can affect an association test is described below (adapted from Thomson and Bodmer 1979): Two bi-allelic markers 1 (A/a) and 2 (B/b) are to be tested in the region of a bi-allelic functional site (X/x). Theoretically, there should be $2^3 = 8$ three-locus haplotypes. In a population, only four three-locus haplotypes exist in equal frequencies: A-x-B (25%), A-X-b (25%), a-x-b (25%), and a-X-B (25%). In this population, the allele X at the functional site has an allele frequency of 50%. The frequency of allele X conditional under an allele at locus 1 or 2 is still 50%. However, while looking at alleles at loci 1 and 2 simultaneously, the allele at the functional site can be perfectly predicted. Nielsen et al. (2004) discussed how three-dimensional LD measures influence haplotype-based association tests and concluded that the multi-locus LD coefficients potentially allow a haplotype-based test to be "greater than the sum of its parts". They found, however, that differences regarding the power to detect associations based on single marker tests vs. haplotype based tests were rather low. Together with the fact that the three-dimensional LD examined in our germplasm set for the SNPs did not reveal an obviously interpretable pattern, we did not combine SNPs to haplotype blocks and also did not perform haplotype based AM.

Association mapping models to analyze data routinely collected in plant breeding programs

As discussed before, population structure can result in spurious associations between markers and phenotypes and, thus, has constrained the use

of AM in human as well as plant research (Flint-Garcia et al. 2003). Recently, Yu et al. (2006) described a mixed model approach (QK model), which allows to account for population structure and relatedness in association analysis. With this approach, the control of both type I and II error rates was improved compared to earlier methods. However, several different methods remain possible, differing in how to calculate the **Q** and **K** matrices. We examined association models which differed with respect to the way of (i) accounting for population structure and (ii) considering pairwise relatedness.

For assessing the population structure matrix **Q**, two methods are frequently used: a Bayesian approach using the software STRUCTURE (Pritchard et al. 2000) or a principal coordinate approach. We found in the Limagrain data set that the two approaches were equally appropriate to assess population structure and, thus, the **Q** matrix from STRUCTURE calculated from the SSR marker data was used for the association analysis.

Several methods for estimating the kinship matrix **K** from molecular marker data have been described to take into account the pairwise relatedness. They are based on the unbiased estimator of the relatedness described by Lynch (1988). Bernardo et al. (1993) extended the model of Lynch (1988) and proposed to calculate the kinship coefficient between inbred A and B (*i.e.*, the probability that inbred A and B carry alleles at the same locus that are identical by descent) on the basis of marker data according to:

$$K_{AB} = \frac{S_{AB} - 1}{1 - T_{AB}} + 1, \tag{3}$$

where S_{AB} is the proportion of marker loci with shared variants between inbred A and B, and T_{AB} is the average probability that a variant from one parent of inbred A and a variant from one parent of inbred B are alike in state, given that they are not identical by descent. In practice, the value of T_{AB} is unknown. The different methods used in the literature to calculate the kinship matrix differ in their way of estimating T_{AB}. Lynch (1988) and Melchinger et al. (1991) proposed to use one T value for all pairs of inbreds

obtained as the average S_{AB} between two sets of unrelated genotypes. Ritland (1996) described a method of moment estimators of T according to the approach proposed by Lynch (1988). Bernardo (1993) calculated an estimate of T_{AB} with

$$T_{AB} = \frac{T_{A.} + T_{B.}}{2}, \tag{4}$$

where $T_{A.}$ (or $T_{B.}$) is the average proportion of variants shared between inbred A (or inbred B) and the unrelated lines. Stich et al. (2008a) proposed a method to obtain a REML estimate of T, compared it with the one from Bernardo et al. (1993), and found an advantage for the former in the context of AM. Contrary to the findings of Stich et al. (2008a), a large deviance of the observed P value distribution compared to the uniform distribution was observed for this QK_T method in the Limagrain data set. However, an incorrect estimation of the optimal T for this data set might explain this observation. The REML calculations for the estimation of the optimal T didn't converge for all T values and, thus, resulted in an incorrect optimal T value.

In the way of estimating the kinship as described by Lynch (1988), a distinction is made between identical in state and identical by descent, where the former is, loosely speaking, defined as the amount of allele sharing expected between unrelated individuals. According to Zhao et al. (2007), this definition would seem to make less sense in the context of AM with elite breeding material, where there are no true unrelated individuals. In that case, it seems to be more promising to define identity by descent as alike in state (Zhao et al. 2007) and to estimate kinship simply as the fraction of shared allele (*i.e.*, $T_{AB} = 0$). With **K** calculated in this way, a lower mean of squared difference between observed and expected P values for the SSR markers was observed compared to the other models. Therefore, this **K** matrix was used in the association analysis.

The QK approach of Yu et al. (2006) includes two terms which target the same aspect. Population structure is here modeled both as fixed effect (by the **Q** matrix) and as random effect (by the **K** matrix). While it is

obvious that **K** can capture the different levels of relatedness that cannot be contained in **Q**, it is not obvious what the **Q** matrix could add in addition to the kinship matrix **K** (Zhao et al. 2007). Therefore, we also examined for both marker types the K alone approach, which revealed to be slightly inferior (higher mean squared difference between observed and expected P values, calculated as a deviation from the uniform distribution) compared to the QK approach. Thus, the QK method described by Yu et al. (2006), with the kinship matrix **K** estimated as described by Zhao et al. (2007) was used for association analysis in the frame of this thesis.

Association mapping of flowering time and northern corn leaf blight resistance and their use by plant breeders

The single marker analyses performed for 8 244 SNP markers revealed seven, four, and four SNP markers significantly ($\alpha = 0.05$, amplicon wise Bonferroni correction) associated with FT, NCLB, and NCLB corrected for FT (NCLB_{FT}) resistance, respectively. These markers explained individually between 0.36 and 14.29% of the genetic variance of the corresponding trait. The very well interpretable pattern of SNP associations observed for FT suggested that data from practical plant breeding programs can be used to dissect polygenic traits. This in turn indicates that the associations identified for NCLB resistance might be successfully used in marker-assisted selection programs. However, when performing the AM in the different heterotic pools, SNPs significantly associated with NCLB and NCLB_{FT} resistance were found in the SSS and Iodent pools, but not in the Flint and Lancaster pools. One explanation could be the difference in the extent of LD between the heterotic pools as discussed before. LD decays more rapidly in the Flint and Lancaster pools compared to the two other pools, and, thus, the number of markers required to detect associations which explain a significant part of

the phenotypic variation is for the Lancaster and Flint pools considerably higher (17 000 and 65 000; respectively) than the number of SNPs we used in the analysis. This could limit the power to detect associations for NCLB and $NCLB_{FT}$ resistance in these two pools. Another explanation could be that the variation in this trait is controlled by a higher number of genes in these two pools, compared to the Iodent and SSS pools.

Alternative approaches to AM: joint linkage association mapping

The inability to identify significant marker-trait associations in our Flint and Lancaster pools could be, as mentioned before, due to the lower power of AM as a consequence of the rapid LD decay in these two pools. Compared to AM, the detection power is higher in bi-parental QTL studies because of the absence of population structure and of low allele frequencies (Yu et al. 2008). However, the mapping resolution in such a population is very low and, in contrast to AM, only two alleles can be studied simultaneously. To combine the advantages of the two methods, joint linkage and association mapping (JLAM) was developed in cattle (Meuwissen et al. 2002). In maize, McMullen et al. (2009) described the nested association mapping (NAM) population that was developed to fulfill this need. The NAM population proved to be very effective to unravel the genetic architecture of complex traits (Buckler et al. 2009; Tian et al. 2011; Kump et al. 2011; Li et al. 2011; Poland et al. 2011). However, this population does not help to unravel the genetic architecture of NCLB in Flint (and Lancaster) germplasm, because it has been developed mainly from Dent inbred lines.

Alternatively, other designs than that used for the NAM population could be used to create JLAM populations. Stich (2009) compared several crossing designs and concluded that for maize, the diallel and factorial crossing

schemes result in the highest power for QTL detection. Such designs, especially the factorial design, which is partly used already in commercial breeding programs for developing breeding populations, could be easily used in maize breeding for QTL detection for NCLB resistance. Even closer to the reality of plant breeder would be association mapping in multiple segregating populations (AMMSP). Contrarily to JLAM, which requires the establishment of segregating populations derived from several crosses of parental inbreds in a systematic manner, AMMSP would use the individuals routinely derived from multiple, related crosses in plant breeding programs (Stich et al. 2008b). Stich et al. (2008b) examined the feasibility of performing AMMSP in elite sugar beet germplasm and found significant marker trait associations. They concluded that AMMSP performed with segregating populations from breeding programs is an interesting alternative to NAM.

Marker-assisted selection programs with markers identified by AM and genomic selection

After identifying marker-trait associations, independently of the method used, the next step for the plant breeder is to use the identified markers in their breeding programs. This process is commonly called marker assisted selection (MAS). MAS is the evaluation of the genetic merit of an individual from a combination of phenotypic value and QTL information (*i.e.*, markers that represent QTL or are linked to QTL). This method proved to work well for traits that meet two criteria (Bernardo 2002): mono- or oligogenic inheritance with a few loci with large effects and difficult phenotypic evaluation (low h^2). However, as soon as the variation in the trait under study is influenced by many QTL with small effects, the published empirical results suggest that MAS is less efficient. This can be explained by the high number of statistical tests that has to be performed in that case, which implies the

use of very stringent significance thresholds. This, as well as the low percentage of variance explained by the QTL, implies in turn a low power of detection and unreliable estimates of the effects. For such cases, there could be an alternative that is not involving the detection of QTL. This approach estimates the effects of all genome-wide markers and uses directly those estimates of the effects to predict the genomic breeding value (GEBV) of the germplasm under consideration (Meuwissen et al. 2001). This method using genome-wide prediction is called genomic selection (GS).

The potential advantage of this method compared to MAS is that having markers covering the entire genome and getting rid of the very stringent significance tests, 100% of the genetic variance could be theoretically explained by the markers (Goddard and Hayes 2007). In contrast to the QTL detected by AM that could help unraveling the genetic architecture of agricultural traits, the GEBVs produced by GS do not give any information of the function of the underlying genes, but are an ideal selection criteria for breeders (Janninck et al. 2010). With the decreasing genotyping costs and the availability of arrays with several hundred thousands of SNPs, as well as the stagnant or even increasing phenotyping costs, GS has revolutionize animal (Schaffer 2006) and plant breeding (Lorenzana and Bernardo 2009).

However, some obstacles had first to be overcome to use GS. The fact that no marker selection is performed by GS implies that a high number of effects has to be estimated (for example for 50 000 SNPs) based on a small number of observations (for instance 2 000 records). With such a shortage of degrees of freedom, standard multiple linear regression cannot be used (Meuwissen et al. 2001). To overcome this problem, various methods like, among others, best linear unbiased prediction (Kolbehdari et al. 2007), ridge regression (Hoerl et al. 2000), and Bayesian regression (Meuwissen et al. 2001) can be used. Therefore, GS could be an interesting alternative for highly polygenic traits, like yield. This method could be used to examine based on NCLB resistance in the Flint and Lancaster pools whether it is the limiting number of available markers that do not allow a marker based genetic improvement

(MAS or GS) or the applied strong significance threshold. In the latter case, this might suggest that the genetic architecture of NCLB in Flint and Lancaster pools is more complex than in the Iodent and SSS pools.

Conclusions

Association mapping methods require background markers to account for genetic relatedness in order to reduce false positive marker-trait associations. The results of this thesis show that both SSR and SNP markers are suitable for this purpose. However, fewer SSRs as SNPs are required to uncover population structure, which facilitates the computations, for instance by the STRUCTURE software. Nevertheless, the findings also indicated that under the assumption of a fixed budget for genotyping and realistic costs for SSR and SNP data points, MRD and D can be more accurately estimated with SNPs than with SSRs. The extent of LD in the study suggested that the 60K SNPs array, currently available for maize, seems appropriate to identify QTL that explain at least 10% of the phenotypic variance. This marker density is definitely needed in the Flint and Lancaster pools, which have a rapid decay of LD. However, to identify QTL with smaller effects, which is a realistic situation for most traits of interest to maize breeders, a much higher marker density is required. The marker-trait associations we identified for FT and NCLB resistance suggested that data from practical plant breeding programs can be used to dissect polygenic traits. The observation that the listed SNPs and their epistatic interactions explained in the entire germplasm set about 10% and in two of the individual heterotic pools up to 30% of the genetic variance suggested that significant progress towards improving the resistance of maize against NCLB by marker-assisted selection is possible with these markers, without compromising on late flowering time. Furthermore, the associated genes are also of interest for further research concerning the mechanism of resistance to NCLB and plant diseases in general, because some of the associated genes have not been mentioned in this context so far.

However, for heterotic pools with a rapid LD decay and for plant breeders aiming to select the best individuals and not to unravel the genes underlying the trait, GS is a promising alternative to select for polygenic traits.

References

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7:781–791

Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet 85:1055–1062

Bernardo R (2002) Breeding for quantitative traits in plants. Stemma Press, Woodbury, p 311, p 319

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Sanchez Villeda H, da Silva HS, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science 325: 714–718

Ersoz ES, Yu J, Buckler ES (2007) Application of linkage disequilibrium and association mapping in crop plants. Genomics-assisted Crop Improvement: Vol. 1: Genomics Approaches and Platforms, 97–119

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124:323–330

Hagdorn S, Lamkey KR, Frisch M, Guimaraes PEO, Melchinger AE (2003) Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations. Crop Sci 43:474–482

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE 2(12):e1367

Hardy OJ, Vekemans X (2002) SPAGedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54–78

Hoerl AE, Kennard RW (2000) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 42:80–86

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics 9:166–177

Jansen RC, Jannink JL, Beavis WD (2003) Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. Crop Sci 43:829–834

Jones ES, Sullivan H, Bhattramakki D, Smith JSC (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (Zea mays L.). Theor Appl Genet 115:361–371

Kolbehdari D, Schaeffer LR, Robinson JA (2007) Estimation of genome-wide haplotype effects in half-sib designs. J Anim Breed Genet 124:356–361

Kump KL, Bradbury PJ, Buckler ES, Belcher AR, Oropeza-Rosas M, Wisser RJ, Zwonitzer JC, Kresovich S, McMullen MD, Ware D, Balint-Kurti PJ, Holland JB (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nature Genetics 43:163–168

Li H, Bradbury P, Ersoz E, Buckler ES, Wang J (2011) Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. PLoS ONE 6(3):e17573

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. Mol Biol Evol 5:584–599

Lorenzana R, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

McMullen MD, Kresovich S, Sanchez Villeda H, Bradbuy P, Li H, Sun Qi, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Oropeza Rosas M, Solvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

Melchinger AE, Messmer MM, Lee M, Woodman WL, Lamkey KR (1991) Diversity and relationships among US maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci 31:669–678

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29

Meuwissen THE, Karlsen A, Lien S, Olsaker I, Goddard ME (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373–379

Nielsen DM, Ehm MG, Zaykin DV, Weir BS (2004) Effect of Two- and Three-Locus Linkage Disequilibrium on the Power to Detect Marker/Phenotype Associations. Genetics 168:1029–1040

Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proc Natl Acad Sci USA 108:6893–6898

Pritchard, JK, Stephens S, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rafalski A (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci 98:11479–11484

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical Research 67:175–185

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123:218–223

Schrag TA, Maurer HP, Melchinger AE, Piepho HP, Peleman J, Frisch M (2007) Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. Theor Appl Genet 114:1345–1355

Stich B, Möhring J, Piepho HP, Heckenberger M, Buckler ES, Melchinger AE (2008a) Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754

Stich B, Melchinger AE, Heckenberger M, Möhring J, Schechert A, Piepho HP (2008b) Association mapping in multiple segregating populations of sugar beet (*Beta vulgaris* L.). Theor Appl Genet 117:1167–1179

Stich B (2009) Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. Genetics 183:1525–1534

Thomson G, Bodmer W (1979) HLA haplotype association with disease. Tissue Antigens 13:91–102

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen DM, Buckler ES (2001) *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet 28:286–289

Tian F, Bradbury PJ, Brown PJ, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011) Genome-wide association study of maize identifies genes affecting leaf architecture. Nat Genet 43:159–162

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Wright S (1938) Size of population and breeding structure in relation to evolution. Science 87:430–431

Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19:395–420

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 2:203–208

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Yu J, Zhang Z, Zhu C, Tabanao DA, Pressoir G, Tuinstra MR, Kresovich S, Todhunter RJ, Buckler ES (2009) Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Genome 2:63–77

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An *Arabidopsis*

example of association mapping in structured samples. PLoS Genet 3:71–82

6. Summary

Besides linkage mapping, association mapping (AM) has become a powerful complement for understanding the genetic basis of complex traits (Yu et al. 2008). AM utilizes the natural genetic diversity and the linkage disequilibrium (LD) present in a diverse germplasm set. *Setosphaeria turcica* is a fungal pathogen that causes northern corn leaf blight (NCLB) in maize. The objective of this thesis research was to set the stage for and perform AM in elite maize breeding populations for NCLB resistance.

Information about the genetic diversity and population structure in elite breeding material is of fundamental importance for the improvement of crops. The objectives of my study were to (i) examine the population structure and the genetic diversity in elite maize germplasm based on simple sequence repeat (SSR) markers, (ii) compare these results with those obtained from single nucleotide polymorphism (SNP) markers, and (iii) compare the coancestry coefficient calculated from pedigree records with genetic distance estimates calculated from SSR and SNP markers. The study was based on 1 537 elite maize inbred lines genotyped with 359 SSR and 8 244 SNP markers.

My results indicated that both SSR and SNP markers are suitable for uncovering population structure. The same conclusions regarding the structure and the diversity of heterotic pools can be drawn from both markers types. However, fewer SSRs as SNPs are required for this goal, which facilitates the computations, for instance by the STRUCTURE software. Finally, the findings indicated that under the assumption of a fixed budget, modified Roger's

distances and gene diversity could be more precisely estimated with SNPs than with SSRs, and we proposed that between 7 and 11 times more SNPs than SSRs should be used for analyzing population structure and genetic diversity.

Association mapping is based on LD shaped by historical recombinations. Many factors affect LD and, therefore, it must be determined empirically in the germplasm under investigation to examine the prospects of genome-wide association mapping studies. The objectives of my study were to (i) examine the extent of LD with SSR and SNP markers in 1 537 commercial maize inbred lines belonging to four heterotic pools, (ii) compare the LD patterns determined by these two marker types, (iii) evaluate the number of SNP markers needed to perform genome-wide association analyses, and (iv) investigate temporal trends of LD.

The results suggested that SNP markers of the examined density, unlike SSR markers, can be used effectively for association studies in commercial maize germplasm. Based on the decay of LD in the various heterotic pools, between 4 000 and 65 000 SNP markers would be needed to detect with a reasonable power associations with rather large quantitative trait loci (QTL). The 60 K SNP chip currently available for maize seems appropriate to identify QTLs that explain at least 10% of the phenotypic variance. However, to identify QTLs with smaller effects, which is a realistic situation for most traits of interest to maize breeders, a much higher marker density is required.

NCLB is a serious foliar disease in maize. In order to unravel the genetic architecture of the resistance against this disease, a vast association mapping panel comprising 1 487 European maize inbred lines was used to (i) identify chromosomal regions affecting flowering time (FT) and NCLB resistance, (ii) examine the epistatic interactions of the identified chromosomal regions with the genetic background on an individual molecular marker basis, and (iii) dissect the correlation between NCLB resistance and FT.

We observed for FT, a trait for which already various genetic analyses have been performed in maize, a very well interpretable pattern of SNP

associations, suggesting that data from practical plant breeding programs can be used to dissect polygenic traits. Furthermore, we described SNPs associated with NCLB and NCLB corrected for FT resistance that are located in genes for which a direct link to the trait is discernable or which are located in bins of the maize genome for which previously QTLs have been reported. Some of the SNPs showed significant epistatic interactions with markers from the genetic background. The observation that the listed SNPs and their epistatic interactions explained in the entire germplasm set about 10% and in some individual heterotic pools up to 30% of the genetic variance suggests that significant progress towards improving the resistance of maize against NCLB by marker-assisted selection is possible with these markers, without much compromising on late flowering time. Furthermore, these regions are interesting for further research to understand the mechanisms of resistance against NCLB and diseases in general, because some of the genes identified have not been annotated so far for these functions.

7. Zusammenfassung

Neben der Kopplungsanalyse hat sich die Assoziationskartierung (AM) als eine vielversprechende Methodenergänzung zur Untersuchung der genetischen Grundlage komplexer Merkmale erwiesen (Yu et al. 2008). Die AM nutzt die natürliche genetische Diversität und das Gametenphasenungleichgewicht (LD), die in einem vielfältigen Genpool bestehen. *Setosphaeria turcica* ist ein pilzlicher Erreger, der die Turcicum-Blattfleckenkrankheit (NCLB) an Mais verursacht. Ziel dieser Doktorarbeit war die Überprüfung der Voraussetzungen einer AM sowie deren Durchführung in Maiselitezüchtungspopulationen für NCLB Resistenz.

Informationen über die genetische Diversität und Populationsstruktur in Elitezüchtungsmaterial sind von grundlegender Bedeutung für die Verbesserung von Kulturpflanzen. Die Ziele dieser Studie waren (i) die Erfassung von Populationsstruktur und genetischer Diversität in Maiselitezüchtungsmaterial anhand von Mikrosatelliten (SSR) Markern, (ii) der Vergleich dieser Ergebnisse mit denen von Einzelbasenpaaraustausch (SNP) Markern, und (iii) der Vergleich des Verwandschaftskoeffizienten berechnet anhand von Abstammungsinformationen mit der genetischen Distanz berechnet mit Hilfe von SSR und SNP Markern. Diese Studie basierte auf 1 537 Maiseliteinzuchtlinien, die mit 359 SSR und 8 244 SNP Markern genotypisiert waren.

Die Ergebnisse dieser Studie zeigten, dass für die Zuordnung der Inzuchtlinien zu Subgruppen mittels SNP Daten und STRUCTURE das Kriterium der höchsten Zugehörigkeitswahrscheinlichkeit angewendet werden

muss, um Subgruppen zu finden, die mit denjenigen, welche anhand von SSR Daten ermittelt wurden, identisch sind. Dennoch können für beide Markertypen die gleichen Schlussfolgerungen bezüglich Populationsstruktur und genetischer Diversität der heterotischen Gruppen gezogen werden. Darüber hinaus zeigten die Ergebnisse, dass unter der Annahme eines festen Budgets modifizierte Roger's Distanzen und die genetische Diversität mit SNP Markern genauer geschätzt werden können als mit SSR Markern. Zusätzlich ergaben die Untersuchungen, dass um ähnlich genaue Schätzwerte der genetische Distanz und Diversität zu erzielen, zwischen 7 und 11 mal mehr SNP als SSR Markern eingesetzt werden müssen.

Die AM nutzt LD, welches durch historische Rekombinationen geformt wurde. Darüber hinaus beeinflussen viele andere populationsgenetische Faktoren das LD. Es ist deshalb erforderlich, das LD in dem interessierenden genetischen Material empirisch zu erfassen, um die Aussichten einer genomweiten AM beurteilen zu können. Die Ziele dieser Studie waren (i) das Ausmaß des LD anhand von SSR und SNP Markern in 1 537 Maiseliteinzuchtlinien aus vier heterotischen Gruppen zu untersuchen und vergleichen, (ii) die Anzahl der SNP Marker, die benötigt werden, um genomweite Assoziationsstudien durchführen zu können, zu bestimmen, und (iii) das Ausmaß vom LD in Inzuchtlinien verschiedener Zulassungsdaten zu vergleichen.

Die Ergebnisse legen nahe, dass die verwendete Zahl von SNP Markern, im Gegensatz zur Zahl der SSR Markern, ausreichend war, um AM in Maiselitezüchtungsmaterial durchführen zu können. Basierend auf der beobachtete Abnahme des LD mit der genetischen Kartendistanz in den verschiedenen heterotischen Gruppen, konnte gezeigt werden, dass zwischen 4 000 und 65 000 SNP Marker benötigt werden, um mit einer angemessenen statistischen Güte Assoziationen mit großen 'Quantitative Trait Loci' (QTL) zu erkennen. Der 60 K SNP Chip, der heutzutage für Mais verfügbar ist, scheint daher notwendig zu sein, um QTL zu erfassen, die mindestens 10% der phänotypischen Varianz erklären. Um jedoch QTL mit kleineren Effekten identifizieren zu können, ist eine wesentlich höhere Markerdichte erforderlich.

NCLB ist ein bedeutende Blattkrankheit von Mais. Mit dem Ziel, die genetische Architektur der Resistenz gegen diese Krankheit zu entschlüsseln, wurden 1 487 europäischen Maiseliteinzuchtlinien zur AM verwendet, um (i) die Genomregionen, die zu Variation des Blühzeitpunktes (FT) und NCLB Resistenz beitragen aufzufinden, (ii) mögliche epistatische Interaktionen der identifizierten Genomregionen mit dem genetischen Hintergrund zu ermitteln, und (iii) die Korrelation zwischen NCLB Resistenz und FT zu untersuchen.

Für FT, für das bereits verschiedene genetische Analysen in Mais durchgeführt wurden, wurde ein sehr gut interpretierbares Muster von SNP Assoziationen beobachtet. Dies belegt, dass Daten aus praktischen Pflanzenzüchtungsprogrammen verwendet werden können, um die polygenen Merkmalen zugrunde liegenden genetischen Faktoren zu detektieren. Darüber hinaus wurden SNP Marker, die assoziiert mit NCLB Resistenz sind, beschrieben, die sich in Genen befinden, für die eine direkte Verbindung zu dem Merkmal erkennbar ist oder sich in Chromosomenregionen des Maisgenoms befinden, in den bereits QTL für dieses Merkmal beschrieben worden sind. Einige der SNP Marker zeigten signifikante epistatische Interaktionen mit Markern aus dem genetischen Hintergrund. Die Beobachtung, dass die ermittelten SNP Marker und deren epistatische Interaktionen im gesamten untersuchten genetischen Material etwa 10% und in einzelnen heterotischen Gruppen bis zu 30% der genetischen Varianz erklärten, legt nahe, dass mit diesen Markern ein beträchtlicher Fortschritt bei der Verbesserung der Resistenz von Mais gegen NCLB durch markergestütze Selektion möglich ist. Darüber hinaus sind diese Regionen interessant für weitere Untersuchungen, um die Mechanismen der Resistenz gegen NCLB sowie andere Krankheiten bei Mais zu verstehen, da einige der identifizierten Gene hiermit bislang noch nicht in Verbindung gebracht worden sind.

8. Acknowledgements

I am very grateful to my academic supervisor Prof. Dr. A.E. Melchinger for his advices, suggestions, and support during this thesis work.

Thanks to Prof. Dr. H.-P. Piepho and Prof. Dr. A. Charcosset for serving on my graduate committee.

Sincere thanks to my academic co-authors, Prof. Dr. B.S. Dhillon and PD Dr. J.C. Reif for proofreading and especially to PD Dr. B. Stich for many interesting discussions and his never ending patience in revising my work.

I would like to thank Mrs. H. Beck and Mrs. S. Meyer for being of great help in organizational matters.

Many thank to my two office mates and two flat mates: Dr. E. Orsini, Dr. C. Von der Ohe, and V. Prigge with whom I shared the good and bad times and enjoyed three years working and living. Many thanks to K. Alheit, S. Bhosale, Dr. C. Bolduan, S. Bopper, PD Dr. C. Falke, Dr. S. Fischer, C. Grieder, M. Hübner, Dr. F. Longin , M. Martin, Dr. H.P. Maurer, PD Dr. H. Parzies, C. Riedelsheimer, Dr. P. Risser, M. Stange, Dr. T. Schrag, D. Schwegler, J. Steinhoff, A. Strigens, F. Technow, Prof. Dr. H.F. Utz, Dr. H.H. Voss, V. Weber, Dr. T. Wegenast, K. Kleinknecht, Dr. B. Müller, Dr. E. Scheuermann; and all unmentioned members of the Institute of Plant Breeding, Seed Science, and Population Genetics for creating a pleasant work environment.

Thanks to Limagrain, the company for which I am working for 12 years, for giving me the opportunity to work on this PhD project. Many thanks to my colleagues Dr. C. Lebreton, P. Flament, and J-P. Martinant who contributed directly to my publications. Furthermore many thanks to A. Blanc,

C. Boyard, S. Chauvet, S. Ducrocq, and Z. Karaman for interesting discussions, and of course to all the breeding teams of Limagrain that were involved in the phenotyping of the inbreds involved in this study.

I want also to thank the members of the Quantitative Crop Genetics group at the Max Planck Institute for Plant Breeding Research for always welcoming me in Cologne when I needed it.

I want to dedicate this work to Mrs. H. Beck and PD Dr. H. Parzies, who unfortunately passed away as I was writing the last sentences of this thesis.

Curriculum vitae

Name	Delphine Laurence Van Inghelandt
Birth	16 August 1975 in Seclin, FRANCE
School education	1981–1986, elementary school (École primaire; Steenvoorde, FRANCE)
	1986–1993, high school (Collège Antoine de Saint Exupery; Steenvoorde, Lycée des Flandres; Hazebrouck, FRANCE) Baccalauréat June 1993
University education	09/93–06/96, Classe préparatoire aux grandes écoles; Lycee Châtelet, Douai, FRANCE
	09/96–10/99, ENSAIA (École nationale supérieure d'agronomie et des industries alimentaires); Nancy, FRANCE Diplome d'ingénieur agronome October 1999
	05/08–04/11, Doctoral Student, Plant Breeding and Applied Genetics, University of Hohenheim; Stuttgart, GERMANY
Agronomy experience	06/97–08/97, Practical experience as farm technician; Angers, FRANCE
	06/98–09/98, Practical experience as institute's technician at RALA, Institute of Agriculture; Reykjavik, ICELAND
Plant breeding experience	03/99–09/99, Thesis work at Tezier; Marmande, FRANCE
	02/00–10/02, Specialty maize breeder for Limagrain Genetics; Aubiat, FRANCE
	11/02–04/08, Maize breeder and station manager for Limagrain GmbH; Greven, GERMANY
	05/11–now, Maize breeder and German research manager for Limagrain GmbH; Schönburg, Pocking, GERMANY

Erklärung

Hiermit erkläre ich an Eides statt, dass die vorliegende Arbeit von mir selbst verfasst und lediglich unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel angefertigt wurde. Wörtlich oder inhaltlich übernommene Stellen wurden als solche gekennzeichnet.

Die vorliegende Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Institution oder Prüfungsbehörde vorgelegt.

Insbesondere erkläre ich, dass ich nicht früher oder gleichzeitig einen Antrag auf Eröffnung eines Promotionsverfahrens unter Vorlage der hier eingereichten Dissertation gestellt habe.

Passau, im September 2011

Delphine Van Inghelandt