Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet: Populationsgenetik
Prof. Dr. Dr. h.c. Hartwig H. Geiger

in Kooperation mit

dem Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung
in Gatersleben
Prof. Dr. Andreas Graner

Bridging genomics and genetic diversity: Association between gene polymorphism and trait variation in a spring barley collection

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

von
Diplom-Agraringenieurin
Grit Haseneyer
geboren in Hennigsdorf

Stuttgart-Hohenheim, 2009

Die vorliegende Arbeit wurde am 12.10.2009 von der Fakultät Agrarwissenschaften als „Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)“ angenommen.

Tag der mündlichen Prüfung: 17.12.2009

1. Prodekan:	Prof. Dr. W. Bessei
Berichterstatter, 1. Prüfer:	Prof. Dr. agr. Dr. h.c. H. H. Geiger
Mitberichterstatter, 2. Prüfer:	Prof. Dr. A. Graner
3. Prüfer:	Prof. Dr. H.-P. Piepho

Contents

---

[1] Haseneyer G, Ravel C, Dardevet M, Balfourier F, Sourdille P, Charmet G, Brunel D, Sauer S, Geiger HH, Graner A, Stracke S (2008) Theor Appl Genet 117:321-331

[2] Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho HP, Graner A, Geiger HH (2009, online first) Plant Breeding DOI: 10.1111/j.1439-0523.2009.01725.x

[3] Haseneyer G, Stracke S, Piepho HP, Sauer S, Geiger HH, Graner A (2010) BMC Plant Biology 10:5

[4] Stracke S, Haseneyer G, Veyrieras JB, Geiger HH, Sauer S, Graner A, Piepho HP (2009) Theor Appl Genet 118:259–273

Abbreviations

AM	–	America
BCC	–	Barley Core Collection
BLZ1	–	Barley leucine zipper 1
BLZ2	–	Barley leucine zipper 2
BPBF	–	Barley prolamin-box binding factor
cM	–	Centi Morgan
DH	–	Double haploide
EA	–	East Asia
EST	–	Expressed sequence tag
EU	–	Europe
HvGAMYB	–	Barley MYB transcription factor
HvCO1	–	Barley homolog to the Arabidopsis *CONSTANS* gene
HvFT1	–	Barley homolog to the Arabidopsis *FLOWERING LOCUS T*
Indel	–	Insertion / deletion event
LD	–	Linkage disequilibrium
MAF	–	Minor allele frequency
MAS	–	Marker-assissted selection
NIL	–	Near-isogenic line
Ppd-H1	–	Photoperiodic response gene
QTL	–	Quantitative trait loci
SNP	–	Single nucleotide polymorphism
SSR	–	Simple sequence repeat
WANA	–	West Asia and North Africa

1. General Introduction

1.1 Association mapping

1.1.1 Definition

Association mapping is defined as the statistical detection and localization of the association between phenotypic trait variation and a polymorphic gene locus in a germplasm collection with different origins and morphological properties (Zhu et al. 2008).

1.1.2 Why association mapping?

Natural phenotypic variation of many agriculturally important traits such as yield or flowering time is the result of the joint action of multiple quantitative trait loci (QTL), environmental effects, and the interaction between the QTL

and the environment (Zhu et al. 2008). The goals of genetic analyses of quantitative traits are the positional cloning of QTL, understanding the molecular nature of quantitative trait variation and discovering genes underlying QTL (Salvi and Tuberosa 2005). Linkage mapping using segregant derived from a biparental cross of contrasting genotypes and association mapping have been frequently used as genetic tools for the analysis of complex traits. The two approaches differ with respect to the size of the mapping population, the extent to which the mapping population represents the target gene pool, and the number and distribution of DNA-markers. QTL positions determined by linkage mapping are less precise and accurate because the confidence interval of the estimated QTL position covers several megabases (Dupuis and Siegmund 1999). Linkage mapping allows only a low genetic resolution, because only one or few meiotic events can occurr between the F_1 and the analysed segregating generation. Moreover, the approach is very costly and needs long time. Turning the gene-tagging efforts from a biparental cross to germplasm collections can reduce the above limitations of linkage mapping.

1.1.3 Concept of association mapping

The term of association mapping is used for two genetic approaches: the genome-wide and the candidate gene approach. Both approaches require large and representative genotype populations (mapping panels), precise phenotypic data for the target trait(s), and multi-environment testing platforms. In the

genome-wide approach populations are genotyped with a large number of genome-wide, evenly distributed molecular markers. To obtain sufficient genomic coverage and thus determine the number of required markers, knowledge about the genome-wide linkage disequilibrium pattern is needed which makes this approach expensive and statistically complex (Hirschhorn and Daly 2005). However, the genome-wide association approach has the potential to discover hitherto unknown QTL and genes contributing to the trait variation. In the candidate gene approach, genotyping is targeted to genes where the annotation and function of the genes underlying the trait are at least hypothetically known and can be used as prior information (Pflieger et al. 2001). Thus, the candidate gene approach builds on genomics resources such as expressed sequence tag (EST) libraries, gene function data of model organisms, and knowledge about the physiology and biochemistry of the trait of interest. The choice of the approach depends on the focus of the particular study.

An association mapping project starts with the composition of a diverse germplasm collection based on available molecular or phenotypic information and passport data (level I). On the second level (level II) the collection is (1) phenotyped for the interesting traits in field trials, or a phenotyping platform or laboratory experiments and (2) genotyped for selected candidate genes and/or with genome-wide distributed markers. Based on a panel of genome-wide evenly distributed markers (3) the population structure is determined and integrated in the statistical association analysis in order to avoid spurious

associations. At the same level (4) pedigree information in form of a kinship matrix is integrated to consider relatedness of genotypes causing in turn population substructure and (5) genome-wide linkage disequilibrium (LD) is estimated. The data resulting from point (1) to (4) are combined in the statistical association model. Genetic effects of the candidate gene's haplotypes and (genome-wide) single nucleotide polymorphisms (SNPs) on trait variation are statistically calculated (level III). Knowing the locus-specific LD allows extraction of markers that are proxy of the functional polymorphism but might be more suitable as diagnostic marker for marker-assisted selection. The identification of diagnostic markers implies progress in molecular plant breeding and cost- and time-saving selection of desired genotypes.

1.2 Linkage disequilibrium

1.2.1 Definition and measurement

Linkage disequilibrium (LD), also known as gametic phase disequilibrium, is the nonrandom distribution of alleles at different genetic loci. It is the correlation between polymorphisms that is caused by their shared history of mutation, selection, and recombination (Flint-Garcia et al. 2003). LD plays the key role in association mapping. The extent of LD in a germplasm collection determines the number and distribution of markers needed to perform association mapping (Nordborg et al. 2005, Yu and Buckler 2006). The

genomic resolution of association mapping is dependent upon the patterns of LD across the genome (Buckler and Thornsberry 2002). Therefore knowledge of LD in a germplasm collection is important to conduct unbiased association mapping (Nordborg et al. 2005).

Many different measures for estimation of LD between two bi- or multi-allelic loci are available. According to Hill (1981), *D* can be calculated as the difference between the product of frequencies of the parental gametic haplotypes and that of the recombinant gametic haplotypes:

$$(1)\ D = q_{AB}q_{ab} - q_{Ab}q_{aB}\,,$$

where q_{AB} is the frequency of the *AB* haplotype in the population, and likewise for the other haplotypes. Values for the *D* statistic range between 0 and 1.

The *D* statistic is very dependent on the frequencies of individual alleles and thus needs standardization to be not suited for comparing the extent of LD among multiple pairs of loci (Hayes, 2006). This is achieved by the r^2 statistic (Hill and Robertson, 1968) which varies between 0 and 1.

$$(2)\ r^2 = \frac{D^2}{q_A q_a * q_B q_b}$$

The statistical significance (*p*-value) of the observed LD is estimated by Monte-Carlo approximation of Fisher's exact test (Weir 1996).

LD can be affected by many genetic and nongenetic factors (Ardlie et al. 2002). Factors which lead to an increase in LD include small effective population size, genetic isolation between lineages, population subdivision, population admixture, natural and artificial selection (Gupta et al. 2005). The mating system also has a profound effect (Gaut and Long 2003). Selfing increases homozygosity, thus decreases the number of double heterozygotes that can be mixed by recombination. As a result, the effective rate of recombination is low in selfing species, genetic polymorphisms tend to remain correlated, and LD is expected to extend over long chromosomal distances (Gaut and Long 2003).

Because LD is highly variable across the genome, it is difficult to obtain a summary statistic of LD across genomes or genomic regions. There are two common ways to visualize LD within a given gene or genomic region: (i) LD decay plots, i.e. the pairwise measures of LD are plotted against the physical or genetic distance between polymorphic sites. They are useful to illustrate the decay of LD along larger physical (several kb) or genetic distances (several cM). (ii) LD matrices, i.e. polymorphic sites plotted on both margins of the matrix and pairwise calculations of LD (e.g. r^2) with corresponding *p*-values are displayed in a heat plot. These heat plots are well suited to display locus-wise LD.

1.2.2 LD in barley

LD studies have been conducted in various plant systems. Knowing the extent of LD is a prerequisite for detecting relationships between nucleotide diversity and phenotypic variation in a populations (Hayes and Szücs 2006). Many LD studies within genes in different barley germplasm collections have been carried out. Results of these studies indicate that LD varies dramatically between different barley loci and germplasm collections. Morrell et al. (2005) examined the LD level within 18 genes of 25 wild barley accessions. They demonstrated that, for the majority of wild barley loci, intralocus LD declines rapidly within 300bp, but a gradual decay is evident above 300bp up to 1,200bp. In an admixed collection of wild barleys, varieties and landraces Caldwell et al. (2006) analysed the hardness locus spanning 212kb and containing four gene loci with regard to inter- and intragenic LD. Intragenic LD indicated high levels of LD extending across the entire gene regions. Intergenic LD values strongly depended on the considered material. In the sample of varieties significant high LD was found across the entire 212kb region. In the landrace sample, significant moderate LD values extended as far as 83kb. Complete equilibrium outside intragenic associations was observed in the wild barleys. Across a collection of 131 accessions, Stracke et al. (2007) found a considerable level of LD occuring within a 132kb physical contig surrounding a locus encoding *Bymovirus* resistance. These pilot studies on LD in barley indicate a strong dependence of LD on the domestication status of the material under study.

1.2.3 Population structure

If a mapping population is structured, i.e. consists of two or more genetically diverse subpopulations, associations between a phenotype and a random marker can arise without physical linkage between the marker and the causal gene (Lander and Schork 1994). Such false positive associations can occur since in structured populations statistically significant LD may occur between completely unlinked loci. Any unlinked marker that is in LD with the causative QTL will then be associated with the phenotype. Such associations are considered spurious, since obviously they are not useful for gene discovery (Oraguzie et al., 2007). Genetic structuring of a mapping population has to be considered in both the genome-wide and the candidate gene-based association approach. Spurious associations if undetected can seriously bias the mapping results. The more distinct the subpopulations are the more inflated will the marker-trait association be if the population structure is neglected. Modern software packages allow the researchers to determine and consider the population structure in association studies (e.g. Pritchard et al., 2000).

1.3 Objectives of the work

The work aims at bridging the areas of genomics and diversity analysis using barley as a model system for a self-pollinated cereal species. With this scope associations were analysed between DNA polymorphisms in selected candidate

genes and variation in flowering time, plant height and grain quality traits. The project comprised four work packages:

1. As a starting point for association mapping and with regard to further association studies a germplasm collection of 224 diverse spring barley accessions was designed as an association platform. The germplasm collection comprises accessions originating from Europe, East Asia, West Asia and North Africa and the Americas. Forty-five expressed sequence tag (EST)-derived simple sequence repeat (SSR) markers were used for the estimation of population structure. The phenotypic evaluation of the collection is conducted in field trials at three locations in 2004 and 2005 in Germany (Haseneyer et al. 2010a).

2. The established association platform was used to investigate the nucleotide and haplotype diversity, and locus specific LD at seven candidate genes to gain further insight in the genetic diversity of the germplasm collection. The photoperiod response gene *Ppd-H1* (Turner et al. 2005), the homologs of the Arabidopsis genes *CONSTANS* (*HvCO1*, Griffiths et al. 2003) and *FLOWERING LOCUS T* (*HvFT*, Faure et al. 2007) as well as four transcription factors, barley MYB transcription factor (*HvGAMYB*), barley leucine zippers 1 and 2 (*BLZ1*, *BLZ2*), and barley prolamin box binding factor (*BPBF*) were investigated.

3. The identification of *Ppd-H1*, *HvCO1* and *HvFT* as players in the regulation of flowering time in barley and other cereals (Distelfeld et al. 2009), and the involvement of *GAMYB*, *BLZ1*, *BLZ2* and *BPBF* in the transcription regulation of B-hordeins encoded by the *Hor2* locus (Gubler et al. 1997; Mena et al. 1998; Onate et al. 1999; Vicente-Carbajosa et al. 1998) makes them attractive as candidate genes for association mapping. With the use of the established association platform associations between nucleotide and haplotype polymorphisms within the seven candidate genes and variation in flowering time (Stracke et al. 2009) and grain quality traits (Haseneyer et al. 2010b) were determined in order to (i) scrutinize the associations of these candidate genes and (ii) identify alleles of these candidate genes to provide tools for practical applications of genomics in barley breeding programs.

4. Nucleotide diversity and LD pattern within *GAMYB* were further examined in a wheat (*Triticum aestivum GAMYB*, *TaGAMYB*) collection to compare polymorphism density and LD pattern between two cereal species (Haseneyer et al. 2008). Location and number of polymorphic sites in the two species indicate to selection pressure in former generations and furnish putative regions for the development of genetic markers useful for MAS for grain quality.

1.4 References

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299-309

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107-111

Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172:557-567

Distelfeld A, Li C, Dubcovsky J (2009) Regulation of flowering in temperate cereals. Curr Opin Plant Biol 12:178-184

Dupuis J, Siegmund D (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373-386

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

Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. Plant Cell 15:1502-1506

Gubler F, Watts RJ, Kalla R, Jacobsen JV (1997) GAMyb: A transcription factor mediating gibberellin-regulated gene expression in aleurone cells of barley. Plant Physiol 114:1493-1493

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461-485

Haseneyer G, Ravel C, Dardevet M, Balfourier F, Sourdille P, Charmet G, Brunel D, Sauer S, Geiger HH, Graner A, Stracke S (2008) High level of conservation between genes coding for the GAMYB transcription factor in

barley (Hordeum vulgare L.) and bread wheat (Triticum aestivum L.) collections. Theor Appl Genet 117:321-331

Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho HP, Graner A, Geiger HH (2010a) Population structure and phenotypic variation of a barley collection set up for association studies. Plant Breeding, DOI: 10.1111/j.1439-0523.2009.01725.x

Haseneyer G, Piepho HP, Sauer S, Stracke S, Geiger HH, Graner A (2010b) DNA polymorphisms and haplotype patterns of transcription factors involved in barley endosperm development are associated with key agronomic traits. BMC Plant Biology 10:5

Hayes P, Szücs P (2006) Disequilibrium and association in barley: thinking outside the glass. P Natl Acad Sci USA 103:18385-18386

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium Genet Res 38: 209-216

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95-108

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037-2048

Mena M, Vicente-Carbajosa J, Schmidt RJ, Carbonero P (1998) An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. Plant J 16:53-62

Morrell PL, Toleno DM, Lundy KE, Clegg MT (2005) Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. P Natl Acad Sci USA 102:2442-2447

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J,

Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196

Onate L, Vicente-Carbajosa J, Lara P, Diaz I, Carbonero P (1999) Barley BLZ2, a seed-specific bZIP protein that interacts with BLZ1 in vivo and activates transcription from the GCN4-like motif of B-hordein promoters in barley endosperm. J Biol Chem 274:9175-9182

Oraguzie NC, Rikkerink EHA, Gardiner SE, de Silva HN (Eds) Association mapping in plants, 2007 Springer Science and Business Media, LLC

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. Mol Breed 7:275-291

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

Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. Trends Plant Sci 10:297-304

Stracke S, Haseneyer G, Veyrieras JB, Geiger HH, Sauer S, Graner A, Piepho HP (2009) Association mapping reveals gene action and interactions in the determination of flowering time in barley. Theor Appl Genet 118:259-273

Stracke S, Presterl T, Stein N, Perovic D, Ordon F, Graner A (2007) Effects of introgression and recombination on haplotype structure and linkage disequilibrium surrounding a locus encoding *Bymovirus* resistance in barley. Genetics 175:805-817

Turner A, Beales J, Faure S, Dunford RP, Laurie DA (2005) The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. Science 310:1031-1034

Vicente-Carbajosa J, Onate L, Lara P, Diaz I, Carbonero P (1998) Barley BLZ1: a bZIP transcriptional activator that interacts with endosperm-specific gene promoters. Plant J 13:629-640

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

Weir BS (1996) Genetic data analysis II, Massachusetts, Sinauer (1996)

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. The Plant Genome 1:5-20

High level of conservation between genes coding for the GAMYB transcription factor in barley (*Hordeum vulgare* L.) and bread wheat (*Triticum aestivum* L.) collections.

Haseneyer G, Ravel C, Dardevet M, Balfourier F, Sourdille P, Charmet G, Brunel D, Sauer S, Geiger HH, Graner A, Stracke S.

Abstract. The transcription factor GAMYB is involved in gibberellin signalling in cereal aleurone cells and in plant developmental processes. Nucleotide diversity of *HvGAMYB* and *TaGAMYB* was investigated in 155 barley (*Hordeum vulgare*) and 42 wheat (*Triticum aestivum*) accessions, respectively. Polymorphisms defined 18 haplotypes in the barley collection and 1, 7 and 3 haplotypes for the A, B, and D genomes of wheat, respectively. We found that (1) *Hv*- and *TaGAMYB* genes have identical structures. (2) Both genes show a high level of nucleotide identity (>95%) in the coding sequences and the distribution of polymorphisms is similar in both collections. At the protein level the functional domain is identical in both species. (3) *GAMYB* genes map to a syntenic position on chromosome 3. *GAMYB* genes are different in both collections with respect to the Tajima D statistic and linkage disequilibrium (LD). A moderate level of LD was observed in the barley collection. In wheat, LD is absolute between polymorphic sites, mostly located in the first intron, while it decays within the gene. Differences in Tajima D values might be due to a lower selection pressure on *HvGAMYB*, compared to its wheat orthologue. Altogether our results provide evidence that there have been only few evolutionary changes in *Hv*- and *TaGAMYB*. This confirms the close relationship between these species and also highlights the functional importance of this transcription factor.

Find full text at www.springerlink.com

Theor Appl Genet. 2008, 117(3): 321-31, DOI 10.1007/s00122-008-0777-4

Population structure and phenotypic variation of a spring barley world collection set up for association studies

Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho HP, Graner A, Geiger HH

Abstract. Association mapping offers a tool to identify plant resources that carry important alleles for crop improvement and breeding. A necessary prerequisite for association mapping is a collection of genotypes representing a cross section of the examined germplasm. This study describes the genetic and phenotypic characterization of a collection of 224 spring barley (Hordeum vulgare L.) accessions sampled from the IPK gene bank. The analysis of the genetic structure of the collection was based on 45 EST-derived simple sequence repeat (SSR) markers and it revealed two major subgroups, mainly comprising two-rowed and six-rowed barleys, respectively. The phenotypic data were based on field trials performed at three locations in Germany in 2004 and 2005. Significant genotypic variation and genotype × environment interaction were observed for all traits under study (thousand-grain weight, crude protein content, starch content, plant height, and flowering time). For all analysed traits entry mean-based heritability estimates exceeded 0.9. After appropriately correcting for population structure and geographic origin significant associations between SSR markers and all traits under study were detected.

Find full text at www3.interscience.wiley.com

Plant Breeding 2009, DOI 10.1111/j.1439-0523.2009.01725.x

DNA polymorphisms and haplotype patterns of transcription factors involved in barley endosperm development are associated with key agronomic traits

Haseneyer G, Stracke S, Piepho HP, Sauer S, Geiger HH, Graner A

BACKGROUND: Association mapping is receiving considerable attention in plant genetics for its potential to fine map quantitative trait loci (QTL), validate candidate genes, and identify alleles of interest. In the present study association mapping in barley (*Hordeum vulgare* L.) is investigated by associating DNA polymorphisms with variation in grain quality traits, plant height, and flowering time to gain further understanding of gene functions involved in the control of these traits. We focused on the four loci *BLZ1*, *BLZ2*, *BPBF* and *HvGAMYB* that play a role in the regulation of B-hordein expression, the major fraction of the barley storage protein. The association was tested in a collection of 224 spring barley accessions using a two-stage mixed model approach.

RESULTS: Within the sequenced fragments of four candidate genes we observed different levels of nucleotide diversity. The effect of selection on the candidate genes was tested by Tajima's D which revealed significant values for *BLZ1*, *BLZ2*, and *BPBF* in the subset of two-rowed barleys. Pair-wise LD estimates between the detected SNPs within each candidate gene revealed different intra-genic linkage patterns. On the basis of a more extensive examination of genomic regions surrounding the four candidate genes we found a sharp decrease of LD (r^2<0.2 within 1 cM) in all but one flanking regions.Significant marker-trait associations between SNP sites within *BLZ1* and flowering time, *BPBF* and crude protein content and *BPBF* and starch content were detected. Most haplotypes occurred at frequencies <0.05 and therefore were rejected from the association analysis. Based on haplotype information, *BPBF* was associated to crude protein

content and starch content, *BLZ2* showed association to thousand-grain weight and *BLZ1* was found to be associated with flowering time and plant height.

CONCLUSIONS: Differences in nucleotide diversity and LD pattern within the candidate genes *BLZ1*, *BLZ2*, *BPBF*, and *HvGAMYB* reflect the impact of selection on the nucleotide sequence of the four candidate loci.Despite significant associations, the analysed candidate genes only explained a minor part of the total genetic variation although they are known to be important factors influencing the expression of seed quality traits. Therefore, we assume that grain quality as well as plant height and flowering time are influenced by many factors each contributing a small part to the expression of the phenotype. A genome-wide association analysis could provide a more comprehensive picture of loci involved in the regulation of grain quality, thousand grain weight and the other agronomic traits that were analyzed in this study. However, despite available high-throughput genotyping arrays the marker density along the barely genome is still insufficient to cover all associations in a whole genome scan. Therefore, the candidate gene-based approach will further play an important role in barley association studies.

Find full text at www.biomedcentral.com
BMC Plant Biology 2010, 10:5, DOI:10.1186/1471-2229-10-5

Association mapping reveals gene action and interactions in the determination of flowering time in barley

Stracke S, Haseneyer G, Veyrieras JB, Geiger HH, Sauer S, Graner A, Piepho HP

Abstract. The interaction between members of a gene network has an important impact on the variation of quantitative traits, and can influence the outcome of phenotype/genotype association studies. Three genes (*Ppd-H1*, *HvCO1*, *HvFT1*) known to play an essential role in the regulation of flowering time under long days in barley were subjected to an analysis of nucleotide diversity in a collection of 220 spring barley accessions. The coding region of *Ppd-H1* was highly diverse, while both *HvCO1* and *HvFT1* showed a rather limited level of diversity. Within all three genes, the extent of linkage disequilibrium was variable, but on average only moderate. *Ppd-H1* is strongly associated with flowering time across four environments, showing a difference of five to ten days between the most extreme haplotypes. The association between flowering time and the variation at *HvFT1* and *HvCO1* was strongly dependent on the haplotype present at *Ppd-H1*. The interaction between *HvCO1* and *Ppd-H1* was statistically significant, but this association disappeared when the analysis was corrected for the geographical origin of the accessions. No association existed between flowering time and allelic variation at *HvFT1*. In contrast to *Ppd-H1*, functional variation at both *HvCO1* and *HvFT1* is limited in cultivated barley.

Find full text at www.springerlink.com

Theor Appl Genet 2009, 118(2):259-73, DOI: 10.1007/s00122-008-0896-y

6. General Discussion

6.1. Generic resource for association mapping in barley

The advantages of population-based association mapping over linkage mapping leads to the most effective application of *ex situ* conserved germplasm resources. The utilization of a sample of individuals of a worldwide collection enables the representation of the natural genetic and phenotypic diversity present in the studied plant species.

In human genetics, the availability of the Human HapMap, a community resource for association mapping describing the common patterns of genetic variation makes studies more efficient and comparable with each other (The-International-HapMap-Consortium 2007). Still existing drawbacks of association mapping in barley, and plants in general, might partly be overcome by the establishment of such community resources. Sequencing candidate

genes in different research groups would rapidly accumulate a large number of characterized alleles for the same germplasm collection. The collaborative analysis of a community resource for association mapping eases studies on LD and promotes the development of tagSNPs, i.e. selected SNPs acting as representatives for the total number of detected SNPs, based on LD. Altogether the availability of a community resource is time- and cost-saving because population structure, phenotypic and genotypic data once recorded can be used for various association mapping studies. Therefore, germplasm collections are being established for genome-wide association mapping in several plant species (rice: Mather et al. 2007, maize: Yu and Buckler 2006, sorghum: Casa et al. 2006; Hamblin et al. 2005). Using the same genotype panels by an entire community of researchers allows to combine and compare result from different studies, a deeper understanding of genetic architecture and mechanism of adaptation, and facilitates the mapping of functional variations (Buckler and Gore 2007).

The establishment of a community resource for spring barley has been started with the development and validation of a worldwide germplasm collection in this study. The collection consists of 224 accessions originating from Europe, America, East Asia, West Asia and North Africa (Haseneyer et al. conditionally accepted). It comprises two-rowed and six-rowed genotypes and reflects various levels of genetic improvement. Phenotypic and genotypic evaluation was conducted to provide a platform for studying associations with agronomically important traits. The collection displays a broad range of

phenotypic variation and high heritabilities were recorded for thousand-grain weight, starch content, crude protein content, plant height, and flowering time (Haseneyer et al., conditionally accepted). Genetic diversity analyses revealed the existence of two genetically distinct subgroups comprising mainly two-rowed and six-rowed spring barleys, respectively (Haseneyer et al., conditionally accepted). Size and composition of the present collection are comparable to association mapping panels in other cereals (Caldwell et al. 2006; Maccaferri et al. 2006). However, in the present collection, due to a high degree of nucleotide diversity at the studied candidate genes many haplotypes (74.2%) occur at frequencies of less than 5%. These haplotypes were excluded from the haplotype-trait association analysis to increase the power of the association analyses. But most of the phenotypic differences were found between those rare haplotypes. A considerably larger germplasm size would be needed to validate those differences. In order to improve the collection for further association studies the collection enlarged by mainly six-rowed, non-European accessions would keep the high degree of diversity and might increase the number of the less frequent haplotypes. Extending the geographical range of origin to non-European accessions would increase the variation of both phenotypes and genotypes (Haseneyer et al., conditionally accepted). Admixing accessions with winter habit would lead to more complicated population stratification (Thiel et al. 2003) so that the advantage of an increased overall diversity might be cancelled by the disadvantage of a reduced within-subgroup diversity. (Haseneyer et al. conditionally accepted).

6.2. Sequence diversity and local LD in the germplasm collection

Nucleotide diversity and LD were investigated in the sequenced fragments (13kb) of seven candidate genes. Within the sequences of those candidate genes (*BLZ1*, *BLZ2*, *BPBF*, *HvCO1*, *HvFT1*, *HvGAMYB*, and *Ppd-H1*) 216 polymorphic sites were detected of which 75 (34.7%) were present at a minor allele frequency (MAF) <0.05. The sequences of *Ppd-H1*, *HvCO1*, and *HvFT* showed a moderate level of nucleotide diversity (Stracke et al. 2009), whereas *HvGAMYB*, *BLZ1*, *BLZ2*, and *BPBF* revealed a high level (Haseneyer et al. submitted). Mean rates between one polymporphic site per 102bp (*Ppd-H1*) and one polymorphic site per 31bp (*HvGAMYB*) were identified. The moderate to high degree of nucleotide diversity observed in the germplasm collection provides the basis for the identification of superior alleles that might be lost by selection in elite germplasm. However, the high level of diversity represented by 224 accessions traces back to the extension of the germplasm collection as discussed in the previous chapter.

Pairwise LD between the detected SNPs revealed different patterns for the genes studied. *Ppd-H1*, *HvCO1*, *BLZ1* and *HvGAMYB* showed strong LD (r^2>0.8, P=0.0001) only between a few polymorphic sites. High LD (r^2>0.8, P=0.0001) was observed across the whole sequence of *BPBF* and *BLZ2* (Haseneyer et al. submitted). A block-like LD structure was detected with r^2>0.8 between sites in the 5'-flanking region of *HvFT* while the mean LD

declined sharply to 0.40 (P=0.0001) in the adjacent coding region (Stracke et al. 2009).

The observed nucleotide diversity and LD in the selected candidate genes may allow a prediction of the number of SNP markers needed to physically cover the entire barley genome in this collection assuming that the estimated LD between polymorphisms within the seven candidate genes is representative for the genome-wide extent of LD. In order to estimate the LD decay along an extended genomic region, additional loci flanking the *BLZ1* gene at increasing distances were investigated. The extended LD study around *BLZ1* revealed a sharp LD decay within 3-5cM and allows a tentative extrapolation. Assuming a genome size of 1,200cM in barley and aiming at three SNPs per cM a number of at least 3,600 SNPs is required to reliably scan the barley genome.

Sequencing of fragments in a germplasm, detecting SNPs and estimating the extent of LD within and between these fragments provide basic insights in the number of markers (e.g. SNPs) that are necessary for genotyping a species with sufficient genome coverage. The chromosomal extent of LD is crucial in the context of genome-wide association mapping because it determines how dense a map must be for detecting significant associations (Nordborg et al. 2005). To date there is no resource available storing genome-wide distributed marker information for barley but sequencing projects are underway (http://barleygenome.org). Genotyping the present collection with markers covering the whole genome would immediately provide a comprehensive view of the genome-wide LD structure. The decreasing costs of genotyping whole

genomes using next generation sequencing technologies will dramatically ease the application of genome-wide association mapping in plants.

As has been shown for the two model plants *Arabidopsis thaliana* and *Oryza sativa*, the availability of an entire genome sequence facilitates map-based gene isolation, which boosts the use of forward genetics approaches to functionally analyse genes involved in natural phenotypic variation of agricultural traits. Furthermore, the availability of a genomic sequence in the association mapping context forms the gateway to study genome-wide patterns of LD (Mather et al. 2007; Nordborg et al. 2002).

The important position of barley within the *Poaceae* family promotes comparisons of gene sequences and genome structure to identify most probable functional regions in the grass genomes. In particular, detailed comparisons to the wheat genome are crucial to reveal conserved structures of chromosomal regions or genes (Feuillet and Keller 1999; Gale and Devos 1998). Sequence and structural homology might explain importance of the selected candidate loci. Differences in such regions or genes launch further investigations to find causal reasons for this variation. In this respect the transcription factor GAMYB was investigated in barley (*Hordeum vulgare*, *HvGAMYB*) and wheat (*Triticum aestivum*, *TaGAMYB*). GAMYB is involved in gibberellin signalling in cereal aleurone cells and in plant developmental processes (Gubler et al. 1999)**.** The high level of conservation observed in *GAMYB* coding sequences in both wheat and barley reflect the importance of this transcription factor in

several developmental mechanisms (Haseneyer et al. 2008). LD pattern and location and number of polymorphisms between species give an idea of the selection history on the selected chromosomal region or gene. A moderate level of LD in *HvGAMYB* was observed whereas in wheat, LD is absolute between intronic polymorphic sites and decays within the gene indicating a lower selection pressure on *HvGAMYB*, compared to its wheat orthologue (Haseneyer et al. 2008) due to a different breeding intensity on the grain protein quality in the two species. In wheat the protein composition is crucial for baking quality whereas in barley only a low total protein content is important for malting barley.

6.3. Association mapping

In plants, missing sequence information, low marker density in the plant genomes, varying LD patterns between germplasm collections of the same species and insufficient homology between species hampers genome-wide association mapping in many plant species. Due to the currently limited number and insufficient density of genetic markers in barley targeted association mapping such as candidate gene-based approaches will continue to play a major role in plant genetics to find marker–trait associations for important traits in barley like grain quality and flowering time.

In the collection described above, association studies on four transcription factors involved in the regulation of B-hordein expression and three candidate

genes playing a role in photoperiodic response were investigated. The four transcription factors *BLZ1*, *BLZ2*, *BPBF*, and *HvGAMYB* were tested for their association to crude protein content, starch content, thousand-grain weight, flowering time and plant height. Haplotypes of *BLZ2* associated with thousand-grain weight, haplotyes of *BPBF* revealed association to crude protein content and starch content. One would expect all four transcription factors associated with the tested grain quality traits. However, the high genetic diversity present in the four loci led to a high percentage of rare haplotypes that were not included in the association analysis. *BLZ1* haplotpyes were associated with flowering time and plant height. Based on the known expression of *BLZ1* in leaves and roots (Vicente-Carbajosa et al. 1998) the observed associations with plant height and flowering time lend strength to the hypothesis that this gene is involved in developmental processes and photoperiodic response. In conclusion, pleiotropic effects of a single gene as observed for *BLZ1* provide a genetic basis and useful information in molecular MAS for multiple traits (Han et al. 1997).

Flowering time belongs to the key traits in agronomy. For instances early flowering is an advantage in regions where the summers are hot and dry (e.g. in West Asia and North Africa) because the plants can complete their life cycle before they are exposed to severe drought. In Central Europe, where the summers are comparatively cool and humid, late flowering is an advantage because the longer growing period allows the crops to deliver higher yields (Hershey 2005). Three genes *Ppd-H1*, *HvCO1* and *HvFT1* known to be

involved in regulation of flowering time (Turner et al. 2005) were selected as candidates for a candidate gene-based association mapping approach (Stracke et al. 2009). Polymorphisms within the candidate genes were significantly associated with flowering time (Stracke et al. 2009). The outcome of this study and results from the literature (Laurie et al. 1994, Laurie 1997) indicate that both linkage mapping and candidate gene-based association mapping are suitable tools for mapping quantitative trait loci influencing grain quality and flowering time.

6.4. Prospects

Verification of the association result

The established collection represents a valuable resource for marker-trait association studies in spring barley (Haseneyer et al. submitted, Stracke et al. 2009). In order to provide favourable alleles for cultivar improvement, the statistically associated candidate gene SNPs or haplotypes might be used in studies for validation (Figure 1). A mandatory task is the investigation of LD around a candidate locus, in order to verify whether in the proximity of the gene there are further genes that contribute to the trait variation. A decay of LD proximal and distal of the candidate gene, as it was observed for *BLZ1*, strengthens the hypothesis that the gene is a causative candidate for the corresponding trait variation. If LD stays high ($r^2 \geq 0.8$) in the vicinity of the candidate gene might indicate that neighbouring genes are involved in the

expression of the trait and selection caused a fixation of alleles in this genomic region.

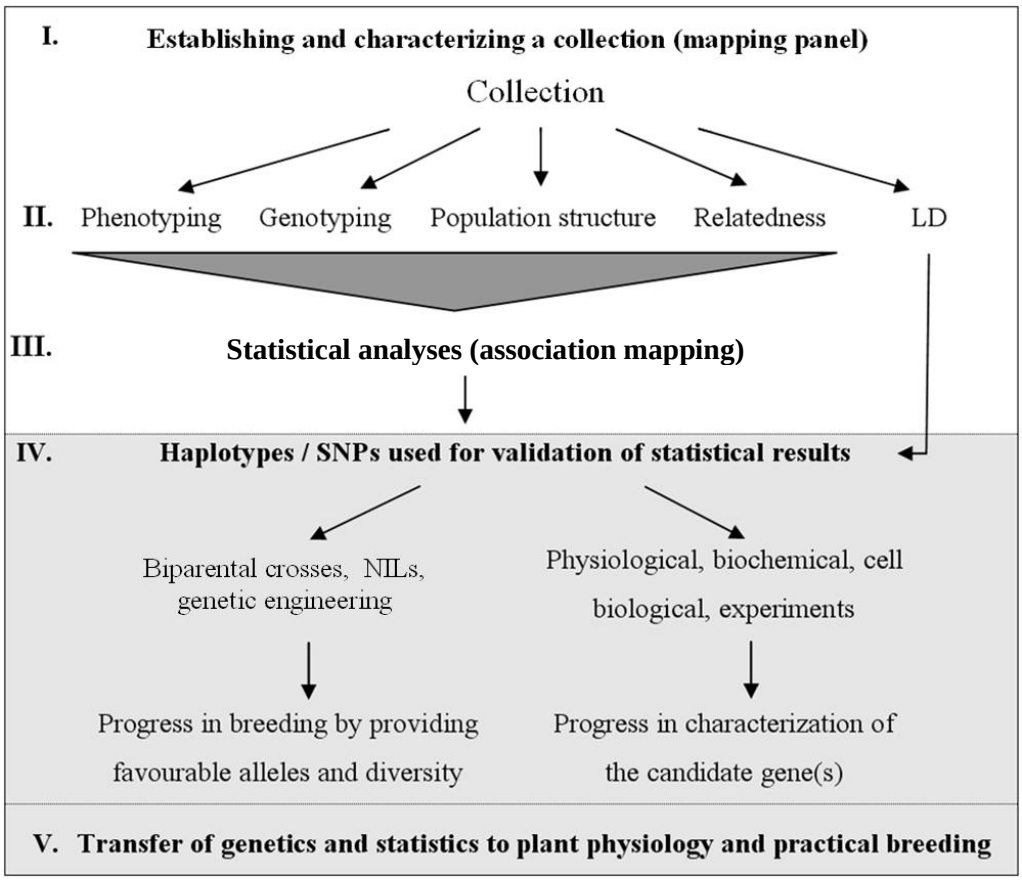

Figure 1. Extended workflow of an association study

The classical workflow of an association analysis is described in chapter 1 (see also Figure 1, level I-III). However, confirmation of candidate gene and genome-wide association results, e.g. in very large populations (Cardon and Palmer 2003), has been proposed in order to reduce the risk of false positive associations. However, even in large association panels, haplotype frequencies might vary among subgroups of a panel (Andersen et al. 2005). Hence, for the development of functional and diagnostic markers in crops that are reliable and

applicable in breeding programs, it might be more useful to test candidate polymorphisms in near isogenic lines (NILs) than to verify associations in large collections (Figure 1, level IV). The possibility of random inter-crossing in plants is a tremendous advantage compared to humans and should be used more efficiently to validate association results. Genetic engineering either by gene silencing or bombardment of the genetic construct containing the associated haplotype sequence might be another possibility to validate the statistical association result. However, the problem of validating single gene effects in the context of quantitative traits might fail due to interrelated gene expression so that effects of one gene can be compensated by others.

It has been learned from several association studies in plants that germplasm collections used for association studies need a broad phenotypic and genetic diversity that is reached by e.g. composing accessions originating from different geographical regions or admixing populations. In addition a sufficient number of individuals per subpopulation is necessary to provide adequate power for the statistical test and to reach a frequency threshold for minor alleles >0.05 for a SNP or haplotpye allele. The high genetic diversity present in the candidate loci especially in *HvGAMYB*, *BPBF*, and *BLZ2* revealed a large number of rare haplotypes (14 out of 18, 15 out of 18, and 18 out of 21). Understanding the phenotypic importance of rare alleles (MAF <0.05) is a critical point for association mapping (Buckler and Gore, 2007). The problem is that only a few genotypes represent the rare allelic state, thus making it impossible to achieve statistical significance (Neale and Sham 2004). If these

rare alleles have important phenotypic effects, then other approaches may be needed. In plants, this issue can be solved by linkage mapping using a bi-parental cross (Figure 1, level IV). Available mapping populations can be screened for the candidate locus and parental genotypes that differ in the respective rare allele can be identified. In the progeny of the cross (DH-lines, F2 generation) more individuals represent the "rare" allelic state and thus phenotypic effects are statistically detectable. By this strategy an advantageous allele, even rarely occurring, can be identified.

Functional examinations of detected associations might be useful to biologically confirm the association results (Figure 1, level IV). The identification of several candidate genes contributing to the genetic variation of an economically interesting trait leads to further scientific hypothesis that entail the establishment of regulatory networks in systems biology. Physiological, biochemical or cell biological experiments might promote the characterization of a candidate gene and its regulation. Both association mapping and confirmation experiments are useful for a better understanding of plant physiological processes and practical application in breeding programmes (Figure 1, level V) e.g. by providing molecular markers for MAS.

If candidate genes have been isolated and confirmed in one species, a possible candidate is available for other crops (Lagercrantz et al. 1996). Therefore, comparisons of polymorphism density and LD pattern between cereal species are interesting not only with respect to evolutionary aspects and selection history but also for the identification, characterization and comparison of

functional alleles in candidate genes. QTL identified by linkage mapping in barley were shown to be informative for predicting putative QTL in wheat (Laurie et al. 1994). For instances, the major photoperiod response locus *Ppd-H1* in barley is located on chromosome 2H (Laurie et al. 1994), which is homoeologous to the group 2 chromosomes of wheat (Law et al. 1978). The studied candidate genes *TaGAMYB* and *HvGAMYB* co-localize with a QTL influencing nitrogen tolerance (Groos et al. 2002; Laperche et al. 2007) and grain protein content (Hayes et al. 1993), respectively. These findings lead to the hypothesis that GAMYB plays a role in nitrogen metabolism and therefore in grain protein content (Haseneyer et al. 2008), particularly as this transcription factor is involved in storage protein synthesis (Diaz et al. 2002).

The transferability of linkage mapping results from one crop to another leads to the question how marker-trait associations in one crop are employable in other related species. Several association studies in plants have been successfully conducted using different candidate genes in various germplasm collections. However, the transferability of association results to different germplasm collections of the same species was not under investigation, except one example. In maize, the general transferability of polymorphisms detected in one collection (Thornsberry et al. 2001) to another collection was investigated (Andersen et al. 2005). The nine polymorphisms within the *Dwarf 8* gene associated with flowering time (Thornberry et al. 2001) were re-analysed in a differently composed association mapping panel. All nine polymorphisms were detected but only two were significantly associated with flowering time.

However, Thornsberry et al. (2001) did not identify associations between plant height and the detected polymorphisms in *Dwarf 8* as Andersen et al. (2005) did. This comparative study indicates that detecting the same polymorphisms in different association mapping panels is possible. However, detected associations for the same trait were not confirmed by the second study so the development of functional markers by association mapping is questionable and traces back to the necessity of validating the association results by linkage mapping or physiological experiments.

Statistical methods

Most statistical approaches for association mapping have been developed in human population genetics. Thus, assumptions for the statistical association model refer to an outcrossing species, a large number of individuals and case- and control populations. An algorithm implemented in the STRUCTURE software for analysing the population structure assumes allele frequencies in Hardy Weinberg equilibrium (Pritchard et al. 2000). This assumption is violated in self-propagating species, the prevalent mating system in plants (Bernardo 2002). However, an extension of the Bayesian STRUCTURE approach eliminates the assumption of Hardy Weinberg equilibrium (Gao et al. 2007). The authors instead calculate expected genotype frequencies on the basis of inbreeding or selfing rates.

Although association mapping in plants takes place on a smaller scale of genotypes (Ching et al. 2002) than in human studies (Gudbjartsson et al. 2008; Lettre et al. 2008; Weedon et al. 2008) the outcome should not be undervalued. Especially phenotypic evaluation in different environments with several replicates provides reliable estimates of the genotypic values for a trait and makes it possible to consider the impact of environmental effects and genotype x environment interactions.

The estimation of associations between sequence polymorphism and trait variation was improved by the development of statistical approaches including factors like population structure, gene action, environment, and genotype by environment interaction in the statistical model. Comparisons of statistical models including different factors and factor combinations showed that the choice of the association model has profound impact on the association result (Casa et al. 2008) especially for minor gene effects below 10% (Zhao et al. 2007). Based on the presented association results the impact of gene interactions on the variation of complex traits is notable (Carlborg and Haley 2004; Hansen et al. 2006).

6.5. Conclusion

The presented work provides further insight in the theory and impact of association mapping in barley. Within the scope of this work a generic resource was established that was shown to be a suitable association platform. The

germplasm collection provides phenotypic data for economically important traits, genotypic data for estimating the population stratification, deep sequencing data of seven candidate genes, and represents a wide range of genotypic and phenotypic diversity. This collection should be enlarged by non-European accessions and may then be introduced as community resource for association mapping in spring barley.

Sequencing of candidate genes creates a redundancy in SNPs with the consequence that in each gene there are SNPs associated and others that are not associated with the considered traits. If only one SNP would be interrogated as is the case with many SNP marker arrays used for whole genome scans there is a risk of overlooking an association in genome-wide association studies as the right SNP was not included in the array. On the other hand, a candidate gene-based approach might lack from the limited knowledge about candidates for a given trait and hence only a part of the genetic variation for this trait is captured. The more knowledge will be accumulated in future about gene function, the better a candidate gene approach will work.

6.6. References

Andersen JR, Schrag T, Melchinger AE, Zein I, Lübberstedt T (2005) Validation of *Dwarf8* polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). Theor Appl Genet 111:206-217

Bernardo, R. 2002. Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN.

Buckler E, Gore M (2007) An Arabidopsis haplotype map takes root. Nat Genet 39:1056-1057

Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172:557-567

Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598-604

Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5:618-625

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30-40

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Diaz I, Vicente-Carbajosa J, Abraham Z, Martinez M, Isabel-La Moneda I, Carbonero P (2002) The GAMYB protein from barley interacts with the DOF transcription factor BPBF and activates endosperm-specific genes during seed development. Plant J 29:453-464

Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. P Natl Acad Sci USA 96:8265-8270

Gale MD, Devos KM (1998) Plant comparative genetics after 10 years. Science 282:656-659

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176: 1635-1651

Groos C, Gay G, Perretant MR, Gervais L, Bernard M, Dedryver F, Charmet G (2002) Study of the relationship between pre-harvest sprouting and grain color by quantitative trait loci analysis in a whitexred grain bread-wheat cross. Theor Appl Genet 104:39-47

Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, Helgadottir A, Ingason A, Steinthorsdottir V, Olafsdottir EJ, Olafsdottir GH, Jonsson T, Borch-Johnsen K, Hansen T, Andersen G, Jorgensen T, Pedersen O, Aben KK, Witjes JA, Swinkels DW, den Heijer M, Franke B, Verbeek AL, Becker DM, Yanek LR, Becker LC, Tryggvadottir L, Rafnar T, Gulcher J, Kiemeney LA, Kong A, Thorsteinsdottir U, Stefansson K (2008) Many sequence variants affecting diversity of adult human height. Nat Genet 40:609-615

Gubler F, Raventos N, Keys M, Watts R, Mundy J, Jacobsen JV (1999) Target genes and regulatory domains of the GAMYB transcriptional activator in cereal aleurone. Plant J 17:1-9

Hamblin MT, Salas Fernandez MG, Casa AM, Mitchell SE, Paterson AH, Kresovich S (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171:1247-1256

Han F, Romagosa I, Ullrich SE, Jones BL, Hayes PM, Wesenberg DM (1997) Molecular marker-assisted selection for malting quality traits in barley. Molecular Breeding 3:427-437

Hansen TF, Alvarez-Castro JM, Carter AJR, Hermisson J, Wagner GP (2006) Evolution of genetic architecture under directional selection. Evolution 60:1523-1536

Haseneyer G, Ravel C, Dardevet M, Balfourier F, Sourdille P, Charmet G, Brunel D, Sauer S, Geiger HH, Graner A, Stracke S (2008) High level of conservation between genes coding for the GAMYB transcription factor in barley (*Hordeum vulgare* L.) and bread wheat (*Triticum aestivum* L.) collections. Theor Appl Genet 117:321-331

Haseneyer G, Stracke S, Paul C, Einfeldt C, Broda A, Piepho HP, Graner A, Geiger HH (conditionally accepted) Population structure and phenotypic variation of a barley collection set up for association studies. Plant Breeding

Haseneyer G, Stracke S, Piepho HP, Sauer S, Geiger HH, Graner A (submitted) DNA polymorphisms and haplotype patterns of transcription factors involved in barley endosperm development are associated with key agronomic traits. BMC Plant Biol

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 germplasm. Theor Appl Genet 87:392-401

Hershey, CH (2005) Plant Breeding News 161 an electronic newsletter of applied plant breeding, Sponsored by FAO and Cornell University; http://www.fao.org/WAICENT/FAOINFO/AGRICULT/AGP/AGPC/doc/services/pbn.html

Lagercrantz U, Putterill J, Coupland G, Lydiate D (1996) Comparative mapping in Arabidopsis and Brassica, fine scale genome collinearity and congruence of genes controlling flowering time. Plant J 9:13-20

Laperche A, Brancourt-Hulmel M, Heumez E, Gardet O, Hanocq E, Devienne-Barret F, Le Gouis J (2007) Using genotype x nitrogen interaction variables to evaluate the QTL involved in wheat tolerance to nitrogen constraints. Theor Appl Genet 115:399-415

Laurie DA, Pratchett N, Bezant JH, Snape JW (1994) Genetic analysis of a photoperiod response gene on the short arm of chromosome 2 (2H) of *Hordeum vulgare* (barley). Heredity 72:619-627

Laurie DA (1997) Comparative genetics of flowering time. Plant Mol Biol 35:167-177

Law CN, Sutka J, Worland AJ (1978) A genetic study of day-length response in wheat. Heredity 41:185-191

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IM, Jacobs KB, Lyssenko V, Uda M, Boehnke M, Chanock SJ, Groop LC, Hu FB, Isomaa B, Kraft P, Peltonen L, Salomaa V, Schlessinger D, Hunter DJ, Hayes RB, Abecasis GR, Wichmann HE, Mohlke KL, Hirschhorn JN (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40:584-591

Maccaferri M, Sanguineti MC, Natoli V, Ortega JLA, Salem MB, Bort J, Chenenaoui C, De Ambrogio E, del Moral LG, De Montis A, El-Ahmed A, Maalouf F, Machlab H, Moragues M, Motawaj J, Nachit M, Nserallah N, Ouabboua H, Royo C, Tuberosa R (2006) A panel of elite accessions of durum wheat (*Triticum durum* Desf.) suitable for association mapping studies. Plant Genetic Resources 4:79-85

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). Genetics 177:2223-2232

Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. Am J Hum Genet 75:353-362

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in Arabidopsis thaliana. Nat Genet 30:190-193

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196

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

Stracke S, Haseneyer G, Veyrieras JB, Geiger HH, Sauer S, Graner A, Piepho HP (2009) Association mapping reveals gene action and interactions in the determination of flowering time in barley. Theor Appl Genet 118:259-273

The-International-HapMap-Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411-422

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

Turner A, Beales J, Faure S, Dunford RP, Laurie DA (2005) The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. Science 310:1031-1034

Vicente-Carbajosa J, Onate L, Lara P, Diaz I, Carbonero P (1998) Barley BLZ1: a bZIP transcriptional activator that interacts with endosperm-specific gene promoters. Plant J 13:629-640

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CN, Morris AD, Ouwehand WH, Zhao JH, Li S, Loos RJ, Barroso I, Deloukas P, Sandhu MS, Wheeler E, Soranzo N, Inouye M, Wareham NJ, Caulfield M, Munroe PB, Hattersley AT, McCarthy MI, Frayling TM (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40:575-583

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

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:e4

7. Summary

Association analysis, initially developed in human genetics, has become common praxis in plant genetics for high-resolution mapping of quantitative trait loci (QTL), validating candidate genes, and identifying important alleles for crop improvement. In the present study the feasibility of association mapping in barley is investigated by associating DNA polymorphisms in selected candidate genes with variation in grain quality traits, plant height, and flowering time to gain further understanding of gene functions involved in the control of these traits.

(1) As a starting point a worldwide collection of spring barley (*Hordeum vulgare* L.) accessions has been established to serve as an association platform

for the present and possible further studies. This collection of 224 accessions, sampled from the IPK genebank, consists of 109 European, 45 West Asian and North African, 40 East Asian and 30 American entries. Forty-five EST derived polymorphic SSR markers were used to determine the genetic structure of the collection. The markers were equally distributed over all seven chromosome pairs. Phenotypic data were assessed in field experiments performed at three locations in 2004 and 2005 in Germany. (2) Seven candidate genes, the barley leucine zippers 1 and 2 (*BLZ1*, *BLZ2*), the barley prolamin box-binding factor (*BPBF*), the barley homologues of the Arabidopsis genes *CONSTANS* (*HvCO1*) and *FLOWERING LOCUS T* (*HvFT1*), the barley MYB factor (*HvGAMYB*), and the photoperiodic response gene *Ppd-H1* were considered. Fragments of these genes were amplified and sequenced in the established collection. Single nucleotide polymorphisms (SNPs), haplotype variants, and linkage disequilibrium (LD) were investigated. (3) The MYB transcription factor was additionally analysed in 42 bread wheat (*Triticum aestivum* L.) accessions in order to compare barley and wheat for nucleotide diversity and LD. (4) Association analysis between SNPs and haplotype variants of the selected candidate genes and the phenotypic variation in thousand-grain weight, crude protein content, starch content, plant height, and flowering time was used to identify candidate genes influencing the variation of these traits in spring barley. A mixed model association-mapping method was employed for this purpose.

In the established collection, significant genotypic variation was observed for all traits under study. Genotype x environment interaction variances were much smaller than the genotypic variances and heritability coefficients exceeded 0.9. Statistical analyses of population stratification revealed two major subgroups, mainly comprising two-rowed and six-rowed accessions, respectively.

Within the sequenced fragments (13kb) of the seven candidate genes, 216 polymorphic sites and 93 haplotypes were detected demonstrating a moderate to high level of nucleotide and haplotype diversity in the germplasm collection. Most haplotypes (74.2%) occurred at a low frequency (<0.05) and therefore were rejected in the candidate gene-based association analysis. Pair-wise LD estimates between the detected SNPs revealed different intra-gene linkage patterns. A high level of LD (r^2>0.8) was observed for *BPBF* and *BLZ2* whereas *Ppd-H1*, *HvCO1*, *BLZ1*, and *HvGAMYB* showed strong LD only between a few polymorphic sites. A block-like LD structure was detected in the 5'-flanking region of *HvFT1* while the mean LD declined sharply to r^2<0.4 in the adjacent coding region. In the flanking region of the *BLZ1* locus LD rapidly decayed to r^2<0.2 within 3 to 5 cM. The 45 SSR markers used for analysing the population structure revealed low intra- and interchromosomal LD (r^2<0.2).

Significant marker-trait associations between the candidate genes and the respective target traits were identified.

The barley and wheat genes *HvGAMYB* and *TaGAMYB*, respectively, showed a high level of nucleotide identity (>95%) in the coding sequences, and the distribution of polymorphisms was also similar in the two species. *HvGAMYB* and *TaGAMYB* both map to a syntenic position on chromosome 3. However, the genes were different in both collections with respect to LD and Tajima's D statistic. In the barley collection only a moderate level of LD was observed whereas in wheat, LD was absolute between polymorphic sites located in the first intron while it decayed by distance between the former sites and those located downstream the first intron. Differences in Tajima's D values indicate a lower selection pressure on *GAMYB* in barley than in wheat.

In conclusion, the established association platform represents an excellent resource for marker-trait association studies. The germplasm collection displays a wide range of genotypic and phenotypic diversity providing phenotypic data for economically important traits and comprehensive information about the nucleotide and haplotype polymorphism of seven candidate genes. Association results demonstrate that the candidate gene-based approach of association mapping is an appropriate tool for characterising gene loci that have a significant impact on plant development and grain quality in spring barley.

8. Zusammenfassung

Die Assoziationsanalyse, entwickelt in der Humangenetik zur Krankheitserkennung, ist auch in der Pflanzengenetik eine aktuelle Methode zur hochauflösenden Kartierung quantitativer Merkmale („quantitative trait loci“, QTL), Validierung von Kandidatengenen und Identifizierung merkmalsrelevanter Allele für die Züchtung. In der vorliegenden Arbeit soll durch die Verknüpfung von Sequenzpolymorphismus in ausgewählten Kandidatengenen und der Variation in den Merkmalen Kornqualität, Pflanzenhöhe und Blühzeitpunkt die Eignung der Assoziationskartierung zur Ermittlung von funktionalen und diagnostischen Markern in Genen, die die Ausprägung dieser Merkmale beeinflussen, geprüft werden.

(1) Für diese und mögliche weitere Assoziationsstudien wurde im Rahmen der Arbeit eine Kollektion ausgewählter Sommergerstenakzessionen (*Hordeum vulgare* L.) der Gaterslebener Genbank zusammengestellt. Die Kollektion umfasst 224 Akzessionen, von denen 109 aus Europa, 45 aus Westasien und Nordafrika, 40 aus Ostasien und 30 aus Amerika stammen. Ein Set von 45 EST(„expressed sequence tag“)-abgeleiteten Mikrosatellitenmarkern („simple sequence repeat“, SSR) wurde zur Bestimmung der genetischen Struktur der Kollektion herangezogen. Die phänotypischen Daten beruhen auf Feldversuchen, die in den Jahren 2004 und 2005 an drei Standorten in Deutschland durchgeführt wurden. (2) Sieben Kandidatengene, *BLZ1*, *BLZ2* („barley leucine zippers“ 1 und 2), *BPBF* („barley prolamine-box binding factor“), *HvGAMYB* („barley MYB factor“), *HvCO1* (Gerstenhomolog des Arabidopsis-Gens *CONSTANS*), *HvFT1* (Gerstenhomolog des Arabidopsis-Gens „*FLOWERING LOCUS T*“) und *Ppd-H1* („photoperiodic response gene“) wurden ausgewählt. Fragmente dieser Kandidatengene wurden amplifiziert und innerhalb der etablierten Kollektion sequenziert. Nukleotiddiversität und Gametenphasenungleichgewicht (Linkage Disequilibrium, LD) wurden untersucht. (3) Der Transkriptionsfaktor GAMYB wurde zusätzlich in einer Kollektion von 42 Brotweizenakzessionen (*Triticum aestivum* L.) hinsichtlich Nukleotiddiversität und LD untersucht. (4) Die Assoziation zwischen den wirtschaftlich bedeutenden Merkmalen Rohproteingehalt, Stärkegehalt, Tausendkorngewicht, Blühzeitpunkt und Pflanzenhöhe und verhältnismäßig häufig vorkommenden Haplotypen bzw. Einzelnukleotidaustauschen („single

nucleotide polymorphisms“, SNPs) der sieben untersuchten Kandidatengene wurde in der oben beschriebenen Kollektion unter Verwendung eines gemischten Modells getestet.

In der etablierten Kollektion konnte eine signifikante genotypische Variation für alle untersuchten Merkmale nachgewiesen werden. Die Varianz der Genotyp x Umwelt Inter-aktion war wesentlich geringer im Vergleich zur genotypischen Varianz, und die Heritabilitätskoeffizienten lagen über 0,9. Statistische Analysen zur Ermittlung der Populationsstruktur ergaben zwei Untergruppen, welche die Akzessionen im Wesentlichen in zweizeilige und sechszeilige Ährentypen unterteilten.

Innerhalb der sequenzierten Fragmente (13kb) der sieben Kandidatengene konnten in der Kollektion 216 SNP-Marker und 93 Haplotypen detektiert werden. Die Mehrzahl der Haplotypen (74,2%) war jedoch nur mit einer geringen Frequenz (<0,05) repräsentiert und wurde deshalb von der Assoziationsanalyse ausgeschlossen. Die LD-Werte zwischen den detektierten SNPs ergaben unterschiedliche Muster für die untersuchten Kandidatengene. Für *BLZ2* und *BPBF* wurde ein hohes Maß an LD (r^2>0,8) beobachtet wohingegen *Ppd-H1*, *HvCO1*, *BLZ1* und *HvGAMYB* nur zwischen wenigen SNPs ein starkes LD zeigten. In der 5’-flankierenden Region von *HvFT1* wurde eine blockähnliche LD Struktur ermittelt, während das LD in dem benachbarten kodierenden Bereich abrupt auf r^2=0,4 abfiel. In der

flankierenden genomischen Region des *BLZ1* Locus' wurde zu beiden Seiten ein starker LD Abbau innerhalb von 3 bis 5cM beobachtet. Das LD zwischen den 45 genomweit verteilten SSR Markern, die zur Ermittlung der Populationsstruktur heranzgezogen wurden, wies ein geringes intra- und interchromosomales LD (r^2<0,2) auf.

In der Assoziationsanalyse konnten signifikante Marker-Merkmal-Assoziationen zwischen den Kandidatengenen und den entsprechenden Merkmalen nachgewiesen werden.

Die vergleichende Analyse der Gerste- bzw. Weizengene *HvGAMYB* und *TaGAMYB* zeigte ein hohes Maß an Nukleotididentität (>95%) innerhalb der kodierenden Sequenz, und auch die Verteilung der Polymorphismen war in beiden Kollektionen ähnlich. *GAMYB* wurde in syntenischen Positionen auf Chromosom 3 kartiert. Im Hinblick auf das LD und auf Tajima's D verhielten sich beide Kollektionen an diesem Locus jedoch unterschiedlich. In der Gerstenkollektion wurde ein mittleres LD beobachtet. Bei Weizen trat zwischen den im ersten Intron lokalisierten Polymorphismen absolutes LD auf, während es zwischen den strangabwärts gelegenen Polymorphismen mit zunehmender Distanz stark zurückging. Diese Unterschiede sind vermutlich auf den, im Vergleich zum orthologen Weizengen, geringeren Selektionsdruck gegenüber *HvGAMYB* zurückzuführen.

Mit der beschriebenen Kollektion wurde eine Ressource geschaffen, die sich hervorragend als Plattform zur Assoziationskartierung bei Sommergerste eignet. Die phänotypischen und genotypischen Daten der Kollektion repräsentieren ein breites Diversitätsspektrum. Die vorliegende Arbeit zeigt, dass die Kandidatengen-basierte Assoziationskartierung eine effektive Methode zur Identifizierung von Genen darstellt, die an der Ausprägung quantitativer Merkmale wie Blühzeitpunkt, Pflanzenhöhe und Kornqualität beteiligt sind.

9. Danksagung

Ich möchte mich ganz herzlich bei Herrn Prof. Hartwig H. Geiger, Herrn Prof. Andreas Graner und Frau Dr. Silke Stracke für die Überlassung des interessanten Themas, die stete Diskussions- und Hilfsbereitschaft, sowie für die Unterstützung bei der Anfertigung dieser Arbeit bedanken.

Bei Herrn Prof. Hans-Peter Piepho möchte ich mich für die intensive Unterstützung bei der statistischen Auswertung und die Übernahme der Aufgabe als Drittprüfer bedanken.

Mein Dank gilt dem Bundesministerium für Bildung und Forschung (BMBF, PTJ-BIO/0313098A), das mich finanziell in dieser Arbeit unterstützt hat.

Ein besonderes Dankeschön möchte ich den Mitarbeitern meiner ehemaligen Arbeitsgruppe Genomdiversität der Abteilung Genbank des Leibniz-Instituts für Pflanzengenetik und Kulturpflanzenforschung für ihre stete Unterstützung bei den molekularen Arbeiten im Labor aussprechen.

Des Weiteren möchte ich mich bei den Züchterhäusern KWS-LOCHOW GMBH und Dr. J. Ackermann & Co. sowie bei Sylwia Koch und Jochen Jesse für die zuverlässige und exzellente Durchführung der zweijährigen Feldversuche bedanken.

Ganz besonders bin ich meinen neuen Kollegen der TU München, Department Pflanzenwissenschaften, Lehrstuhl Pflanzenzüchtung für die freundliche Atmosphäre dankbar. Im Besonderen danke ich Prof. Chris-Carolin Schön für die zugestandene Zeit zur Fertigstellung der Arbeit und die intensive Diskussion über meine Arbeit, Dr. Eva Bauer für das entgegengebrachte Verständnis und Silke Wieckhorst für das Korrekturlesen und ihre Diskussionsbereitschaft zu jeder Zeit.

Von ganzem Herzen möchte ich Dr. Heike Thiel für ihre Geduld, Unterstützung und innige Freundschaft sowie das Korrekturlesen dieser Arbeit und den immerwährenden Glauben an mich, diese Arbeit fertig zu stellen, danken.

Mein tiefer Dank gilt meinen Eltern, für ihre immerwährende Geduld und stete Unterstützung.

10. Curriculum vitae

PERSONAL DATA

Name:	Grit Haseneyer
Date of birth:	11th November, 1978
Place of birth:	Hennigsdorf, Germany

EDUCATION

2008 – present	Research Associate, Technische Universität München
2004 – 2007	PhD, Universität Hohenheim, Stuttgart, Germany and Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany
1998 – 2003	Study of Horticulture Science, Leibniz Universität Hannover Diplomathesis: „Functional and molecular characterization of small cell wall proteins in *Arabidopsis thaliana*“
1998	Abitur, Alexander S. Puschkin Gymnasium, Hennigsdorf, 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

Freising, 18.12.2009

_____

Grit Haseneyer