TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Pflanzenzüchtung

Association analysis of frost tolerance in rye (*Secale cereale* L.) using candidate gene polymorphisms and phenotypic data from controlled, semi-controlled, and field phenotyping platforms

Yongle Li

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Agrawissenschaften (Dr.agr.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. C. Schwechheimer

Prüfer der Dissertation: 1. Univ.-Prof. Dr. C.-C. Schön
2. Univ.-Prof. D. P. Ankerst, Ph.D.

Die Dissertation wurde am 02.01.2012 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 08.03. 2012 angenommen.

Contents

Abbreviations

AFLP	Amplified fragment length polymorphism
AMOVA	Analysis of molecular variance
ANOVA	Analysis of variance
BLAST	Basic local alignment search tool
Cbf	C-repeat binding factor
ChIP	Chromatin immunoprecipitation
CI	Confidence interval
CG	Candidate gene
CDS	Coding sequence
COR	Cold responsive
Dhn	Dehydrin
DHs	Doubled haploids
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
Dreb2	Dehydration responsive element binding gene 2
EST	Expressed sequence tag
FAO	Food and Agriculture Organization of the United Nations
FT	Frost tolerance
GWAS	Genome-wide association study
h^2	Heritability
Hd	Haplotype diversity
Ice2	Inducer of Cbf Expression 2
Indels	Insertions and deletions
KAS	Kasan, Russia
LEA II	Late embryogenesis abundant II gene
LD	Linkage disequilibrium
LIP	Lipezk, Russia
LRT	Likelihood ratio test
LSD	Least significant difference
MAB	Maker-assisted backcrossing
MAF	Minor allele frequency
MAS	Marker-assisted selection
Mb	Megabase
MIN	Minsk, Belarus
$MgCl_2$	Magnesium chloride

NAM	Nested association mapping
PCoA	Principal co-ordinate analysis
PCR	Polymerase chain reaction
PIC	Polymorphism information content
QTL	Quantitative trait locus
RAPD	Random amplified polymorphic DNA
RCBD	Randomized complete block design
RFLPs	Restriction fragment length polymorphisms
RILs	Recombinant inbred lines
RMS	Rye microsatellite (from Khlestkina et al. 2004)
R_s	Allelic richness
SAS	Saskatoon, Canada
SCM	*Secale cereale* microsatellite (from Hackauf and Wehling 2002)
SD	Standard deviation
SNPs	Single nucleotide polymorphisms
SSRs	Simple sequence repeats
UPGMA	Unweighted pair group method with arithmetic mean
Vrn1	Vernalization gene 1

List of figures

List of tables

1. Introduction

Cold or frost stress is one of the important abiotic stresses which not only limits the geographic distribution of crop production but also adversely affects crop development and yield through cold-induced desiccation, cellular damage and inhibition of metabolic reactions (Chinnusamy et al. 2007; Gusta et al. 1997). Cold stress leads to chilling injury at temperatures between 10°C to 0°C whereas frost stress leads to freezing injury at temperatures lower than 0°C. Crop varieties with improved tolerance to frost are of enormous value for Middle and Eastern Europe with severe winters where daily average air temperature in January between 1995 and 2003 dropped down to between 3°C and -18°C (Figure 1). Frost tolerance (FT) is one of the most critical traits that determine winter survival of winter cereals (Saulescu and Braun 2001). Among small grain cereals, rye (*Secale cereale* L.) is the most frost tolerant species and thus can be used as a cereal model for improving FT (Fowler and Limin 1987; Hömmö 1994). After exposure to a period of non-freezing temperature, the so-called cold acclimation, the most frost tolerant rye cultivar can survive severe frost stress down to about -30°C (Thomashow 1999).

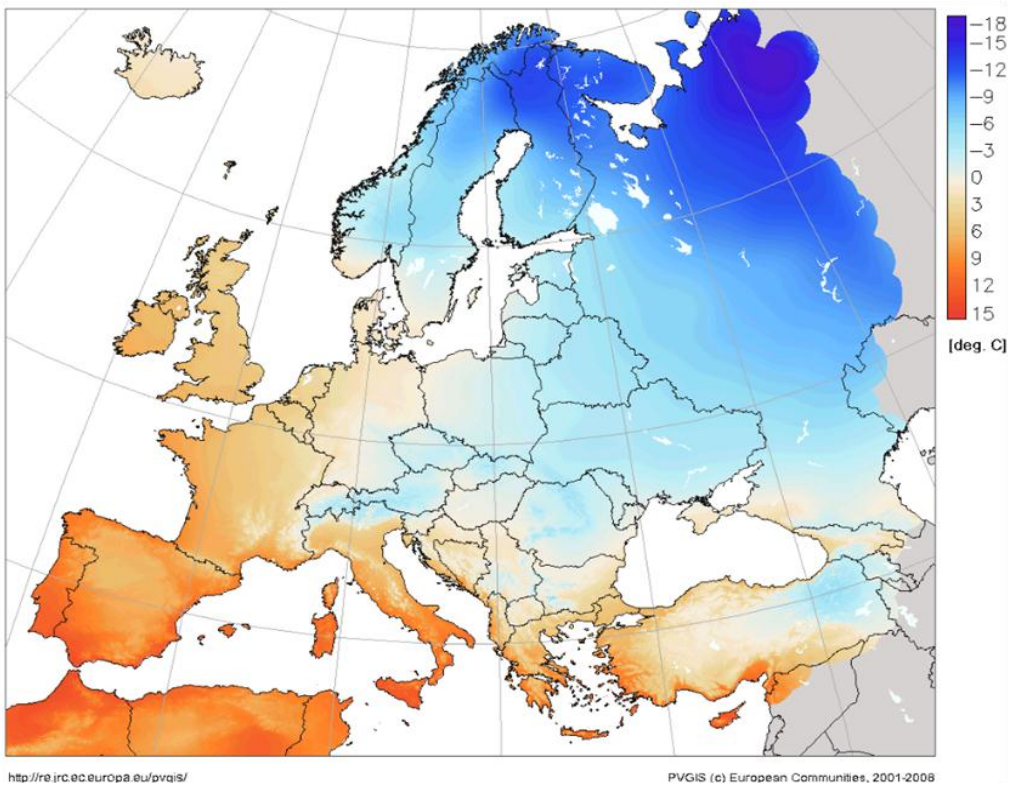


Figure 1 Daily average air temperature in January between 1995-2003 in Europe (Huld et al. 2006).

1.1 Cultivation and uses of rye

Rye is a cross-pollinated cereal with a diploid genome and closely related to wheat and barley. The haploid genome of rye consists of seven chromosomes (2n=14) with a total estimated genome size of ~8,000 Mb, which is more than sixty times larger than the genome size of *Arabidopsis thaliana* (Zonneveld et al. 2005). The domestication process of cultivated rye is still not clear (Shang et al. 2006). However, it is generally accepted that the first cultivation of rye took place in the Fertile Crescent of Middle East and from there spread to Europe 2000-2500 BC (Khush 1963). From the Middle Ages until the beginning of the 20th century rye was a major crop in Middle and Eastern Europe and once even surpassed wheat in cultivation area due to its better performance under unfavourable conditions such as frost, drought, and nutrient-poor soil (Murphy 2007). Despite the fact that rye originated from warm areas (the Fertile Crescent of Middle East), it is nowadays mainly cultivated in the temperate zone with severe winter conditions. Since the civilization moved northward toward Europe, there has been intensive selection for FT in rye. Alleles which were not well adapted to new climatic conditions were lost while mutations that gave a selective advantage against frost were favoured and eventually fixed (Revilla et al. 2005).

According to data from FAO in 2009, rye is the fourth most important cereal in Europe in terms of production (8^{th} in the world, FAOSTAT, 2011). The top ten rye production countries are Russia, Germany, Poland, Belarus, Ukraine, Turkey, China, Canada, Spain, and USA (Figure 2). It is mainly grown in Europe with approximately 1.64×10^{7} tons annually corresponding to 84% of world production in 2009 (FAOSTAT, 2011). Although the harvest area of rye has decreased globally by more than one half since the 1960s, rye is still grown on about 6 million hectares in Europe due to its superior adaptation capability to harsher winter and poor, light soil and hence outperforms wheat and barley in less favourable environments. Besides, rye bread becomes more and more popular as healthy food because of its high soluble fibre and other beneficial nutrients such as phenolic compounds, vitamin B and vitamin E (Bondia-Pons et al. 2009).

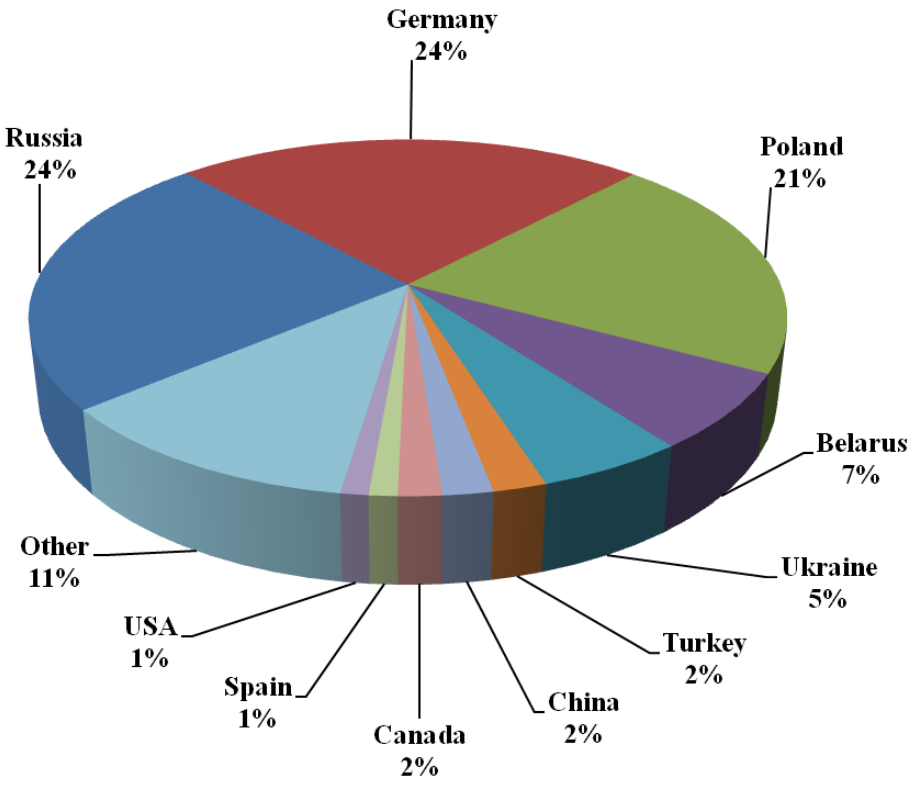


Figure 2 Pie chart of top ten rye production countries in 2009 based on annual harvest quantity (2009 FAOSTAT).

Rye can be used for many purposes. In the growth stage before heading, rye is used as livestock pasture and green manure for crop rotation; in the harvest stage as grain, it is used as livestock feed and for vodka production; in the postharvest stage as flour, it is used for making “black” bread, consumed extensively in Middle and Eastern Europe (Bushuk 2001). Rye straw can be used to produce paper, insulation materials and renewable energy such as bioethanol. Due to these features, rye will remain an important cereal in Middle and Eastern Europe in crop rotation. Germany is the top exporter of rye contributing 39% of the rye export quantity in the world during 2008. The average yields of rye in Germany have been increased ~50% since 1984. However, German breeders are facing at least two challenges: 1) Reduction of genetic diversity in elite material; 2) Development of new elite material adapted to new expanding economic markets, e.g. Eastern Europe with continental climate and severe winters. Thus, there is a constant need to integrate desirable traits such as FT from the well-adapted populations of the geographical target region.

1.2 Genetic diversity of rye

Genetic diversity, the result of mutation and recombination, is vital to crop improvement. DNA markers, typically derived from a small region of DNA that show sequence polymorphisms between individuals, have been an indispensable tool for assessing genetic diversity. Many types of DNA markers have been used for this propose in various crop species such as restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphisms (AFLPs), Diversity Arrays Technology (DArT), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs). Among these, the latter three are particularly valuable because their detection is highly reproducible and can be easily automated (Jones et al. 2009). For example, around two thousand DArT markers distributed across the rye genome with one unique marker every 2.68cM has been developed for assessing genetic diversity and mapping (Bolibok-Bragoszewska et al. 2009). Another advance in rye genomics resource is the development of a Rye5K high-throughput SNP array generated using rye transcriptome sequencing (Haseneyer et al. 2011). With the help of these new technologies, exploiting genetic diversity in a non-model species like rye becomes more and more cost-effective.

Investigating genetic diversity within and between populations is important for developing hybrid rye. Higher genetic variation within than between populations was observed when investigating 26 rye populations from Northern Europe with isoenzyme markers (Persson and von Bothmer 2000) and 12 rye populations from Northern Europe with RAPD markers (Persson et al. 2001). Similar results have also been reported in other outcrossing species, including white clover (George et al. 2006) and perennial ryegrass (Bolaric et al. 2005), probably a consequence of the obligate cross-pollinated reproductive behaviour of outcrossing species (Rafalski 2002). On the contrary, investigations in the self-pollinated species rice have revealed larger variation between populations (Zhang et al. 2009). A wide ecological and geographical distribution of rye probably contributes to its high genetic diversity within different populations. However, elite breeding material will inevitably suffer from reduction of genetic diversity due to genetic drift and high selection pressure. For example, a recent study has been done to investigate the genetic diversity of five Eastern European populations and the two heterotic rye breeding pools Carsten and Petkus (Fischer et al. 2010). Around 600 S_0 rye plants were fingerprinted with 30 genome-wide distributed SSR markers revealing that the heterotic pool Carsten was genetically different from the Petkus pool and had less genetic diversity than the five Eastern European populations. Therefore,

broadening the genetic diversity of core breeding material is one of the most important cornerstones of a successful breeding program.

1.3 Frost tolerance in plants

Precise frost tolerance (FT) phenotyping is time and labour consuming. There are numerous ways to test FT which can be generally separated into direct and indirect approaches. For direct approaches, where plants are exposed to both cold acclimation and freezing tests, plant survival rate, leaf damage, regeneration of the plant crown, electrolyte leakage, and chlorophyll fluorescence are often used as phenotypic endpoints (Saulescu and Braun 2001). For indirect approaches, where plants are exposed to only cold acclimation, end points such as water content (Fowler et al. 1981), prolines (Dorffling et al. 1990), and frost induced proteins (Houde et al. 1992) are commonly used. The evaluation of FT can be conducted either under field conditions or in growth chambers. Both methods are associated with advantages and disadvantages. Under field conditions, plant damage during winter is often not only affected by low temperature stress *per se*, but also by the interaction of a range of factors such as snow coverage, water supply, and wind. Therefore, measured phenotypes are the result of the full range of factors affecting winter survival. Opportunities for assessing FT are highly dependent on temperature and weather conditions during the experiment. In contrast, frost tests in growth chambers allow for control of experimental error and are not limited to one trial per year. However, they are limited in capacity, more expensive to perform, and may not correlate well with field performance. Therefore, it has been recommended to test FT under both natural and controlled conditions whenever possible (Saulescu and Braun 2001).

Investigating the functional genetic base of FT in expression studies performed under controlled conditions showed that FT is a complex trait with polygenic inheritance (Hannah et al. 2005; Kreps et al. 2002; Vogel et al. 2005). Metabolite profiling in *Arabidopsis* has revealed between 311 (63%) and 434 (75%) metabolites altered in response to cold (Cook et al. 2004; Kaplan et al. 2004). Among these, glucose, galactose, fructose, raffinose, sucrose, and xylose are involved in the central carbohydrate metabolism and play a prominent role during reprogramming of metabolism under cold stress. A large number of genes are up- and down-regulated when plants are exposed to cold/frost stress. Transcriptome analyses have estimated between 14% and 45% of the *Arabidopsis* genome to be cold responsive, depending on the cold treatment and other experimental factors (Hannah et al. 2005; Kreps et al. 2002; Vogel et al. 2005). Studies in wheat have also shown between 5% and 8% of

transcripts represented on microarrays to be regulated under cold stress (Monroy et al. 2007; Winfield et al. 2010). These genes are mainly involved in stress signalling, transcriptional regulation, and direct response to cold/frost, including cellular membrane stabilization. A model of the frost responsive network in plants is illustrated in Figure 3 structured in three levels. In level 1, the gene *Inducer of Cbf Expression 2* (*Ice2*) is a basic helix-loop-helix transcription factor that can bind to *cis*-elements (Ice box) in the promoters of the *C-repeat Binding Factor* (*Cbf*) gene family and activate their transcription under frost stress in hexaploid wheat (Badawi et al. 2008). Over-expression of *Arabidopsis Ice2* (Fursova et al. 2009) resulted in increased tolerance to deep freezing stress at a temperature of -20°C after cold acclimation. In level 2, the *Cbf* gene family belongs to the family of *APETALA2* transcription factors. In barley as well as in diploid and hexaploid wheat, several cereal *Cbf* homologs have been cloned and mapped to the *Fr2* locus on homoeologous group 5, which coincides with a major QTL for FT (Baga et al. 2007; Francia et al. 2007; Knox et al. 2008). Using wheat-rye addition lines, Campoli et al. (2009) assigned twelve members of the *Cbf* gene family to the long arm of chromosome 5R in rye. Several studies in *Arabidopsis* provided evidence that allelic variation in the *Cbf* gene family formed the molecular basis for the freezing tolerance QTL (Alonso-Blanco et al. 2005; McKhann et al. 2008). *Cbf* transcription factors activated *Cold Responsive* (*COR*) genes through binding to *cis*-elements (CRT/DRE) in the promoters of *COR* genes under cold stress in *Arabidopsis* (Chinnusamy et al. 2007). More than 70 proteins encoded by *COR* genes were involved in direct response to cold/frost forming level 3 in the frost-responsive network. Dehydrins, also known as Late Embryogenesis Abundant II (LEA II) proteins, were among these proteins that protected other proteins and membranes from cellular damage caused by dehydration (Kosova et al. 2007). In barley, 13 dehydrin genes (*Dhn 1-13*) belonging to the *COR* gene family have been identified (Choi et al. 1999). Transcripts of *Dhn1, Dhn2, Dhn3, Dhn4, Dhn7,* and *Dhn9* were detected in plants subjected to cold acclimation at 4°C followed by mild frost at -2°C or -4°C (Zhu et al. 2000). *Dhn1* and *Dhn3* were mapped in barley to chromosome 5H near a QTL for winter hardiness and on chromosome 6H, respectively (Kosova et al. 2007). Recent studies showed that cold/frost regulation and vernalization were interconnected (Galiba et al. 2009; Kosova et al. 2008). Winter cereals require long exposure to cold in winter, the so-called vernalization, to accelerate flowering in the next spring. This process prevents the early transition of winter cereals into the less cold-tolerant reproductive phase. The vernalization gene *Vrn1* has been mapped to the second locus conferring frost tolerance, *Fr1*, on the long arm of homoeologous group 5 near the *Fr2* locus (Galiba et al. 1995). Transcript levels of all

cold-induced *Cbf* genes at the frost tolerance locus *Fr2* in barley were significantly higher in lines possessing the *vrn1* winter allele than in lines possessing the *Vrn1* spring allele (Stockinger et al. 2007). It remains unknown how the *Cbf* family members interact with *Vrn1* under frost stress. Despite a fairly clear understanding of the frost-responsive network on the gene level, relative little is known about the precise functional polymorphisms that determine FT which are of enormous value in plant breeding. Candidate gene-based association studies, described in the next chapter, are a valuable tool for marker-assisted selection (MAS) in rye breeding.

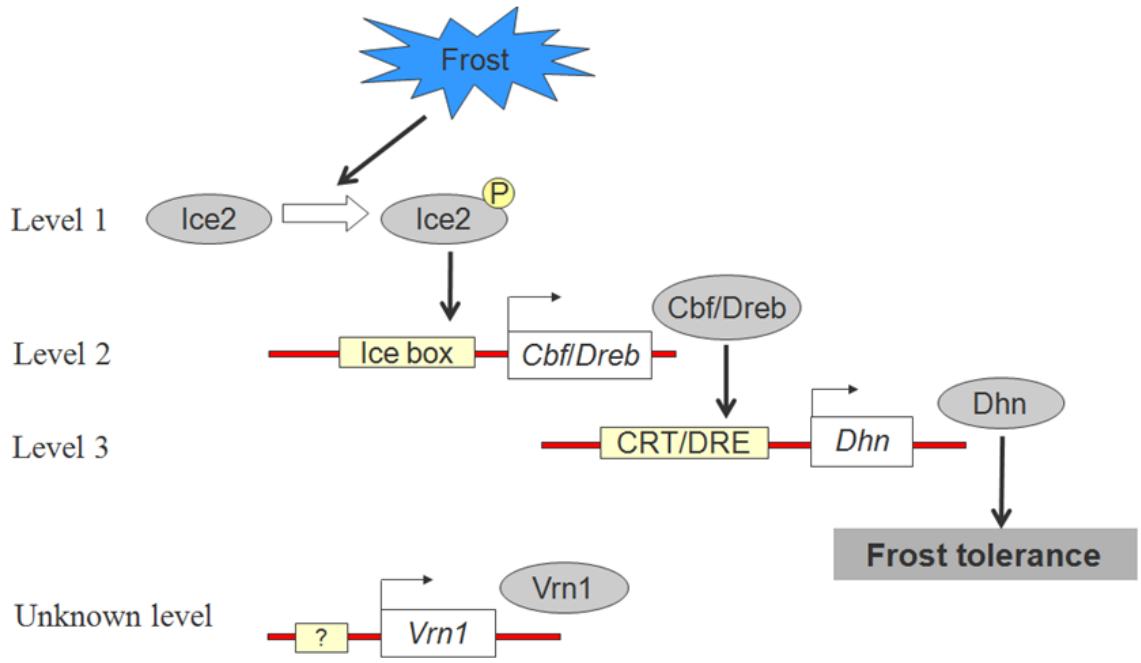


Figure 3 Model of the frost-responsive network in plants (modified after Yamaguchi-Shinozaki and Shinozaki 2006).

Grey ovals illustrate gene products of *Ice2*, *Cbf* gene family, *Dhn* gene family, and *Vrn1*. The small yellow circle indicates the phosphorylation of transcription factor *Ice2* in response to frost stress. Red bars show the DNA strand while yellow and white boxes represent *cis*-acting elements and coding sequences of the genes, respectively. The mechanism that triggers the expression of *Vrn1* under frost stress and the interaction between *Vrn1* and *Cbf* gene family members are still unknown.

1.4 Association studies and linkage disequilibrium

Identification of genes and functional polymorphisms underlying traits of agronomic interest is pivotal for genome-based breeding. Genomics has developed many tools to identify genes that play a role in simply inherited traits. However, relating individual genes and alleles to complex traits such as abiotic stress tolerance is still challenging. Due to methodological

advances in molecular biology, plant breeders can now select varieties with molecular markers, such as SSRs and SNPs (Rafalski 2002; Tester and Langridge 2010). Bi-parental linkage mapping using segregating populations derived from crosses between parents with contrasting phenotypes is an important tool for discovering QTL for plant improvement. However, producing high resolution linkage maps with recombinant inbred lines or near isogenic lines is labour intensive and time consuming. An alternative approach, association studies, using ancestral recombination and natural genetic diversity within populations to study QTL, promises to compensate those shortcomings. At least three advantages can be anticipated by association studies: an increased mapping resolution, an increase in examined alleles and a reduced research time (Yu and Buckler 2006). Association studies generally fall into two categories: genome-wide and candidate gene-based association studies. The former utilise hundreds or thousands of markers covering the whole genome to find signals of association between markers and various complex traits; the latter tries to associate specific traits with markers from selected candidate genes that have putative roles in controlling phenotypic variation of the investigated traits. Association studies have identified a large number of genomic regions and individual genes in plants related to a range of traits (Table 1). However, underlying population structure and/or familial relatedness between genotypes under study have proven to be a big challenge, leading to false positive associations between molecular markers and traits due to the heavily admixed nature of plant populations (Aranzana et al. 2005). In response, several advanced statistical approaches have been developed for genotype-phenotype association studies, including genomic control (Devlin and Roeder 1999), inference of population structure (Pritchard et al. 2000c), and linear mixed model-based methodologies (Stich et al. 2008a; Yu et al. 2006). The latter estimates population structure via a structure matrix and familial relatedness via a kinship matrix in a first step, and then includes these as covariates in a linear mixed model comprising the second step, thus arriving at phenotype-genotype association studies adjusting for population structure and kinship. This approach has proven to be the most effective approach and becomes the method of choice in plant and animal association studies (Kang et al. 2008; Yu et al. 2006). The general workflow of this approach adapted to this study is illustrated in Figure 4.

Table 1 Summary of recent association studies in plants

Plant species	Sample size	Types of populations	Traits	CG or GWAS[b]	Reference
Arabidopsis	95	Diverse accessions	Flowering time	CG	Olsen et al.(2004)
	95	Diverse accessions	Disease resistance	GWAS	Aranzana et al.(2005)
	96	Diverse accessions	Shoot branching	CG	Ehrenreich et al.(2007)
	275	Diverse accessions	Flowering time	CG	Ehrenreich et al.(2009)
	473	Diverse accessions	Flowering time	GWAS	Li et al.(2010b)
	199	Diverse accessions	107 traits	GWAS	Atwell et al.(2010)
Barley	220	Diverse accessions	Flowering time	CG	Stracke et al.(2009)
	224	Diverse accessions		CG	Haseneyer et al.(2010)
	102	Core germplasm	Growth habit	GWAS	Cuesta-Marcos et al.(2010)
Douglas fir	700	Diverse families	Cold hardiness	CG	Eckert et al.(2009)
Maize	92	Inbred lines	Flowering time	CG	Thornsberry et al. (2001)
	553	Inbred lines	Kernel quality	CG	Wilson et al.(2004)
	282	Inbred lines	Carotenoid content	CG	Harjes et al.(2008)
	4892[a]	Inbred lines	Leaf architecture	GWAS	Tian et al.(2011)
Pearl millet	598	Diverse accessions	Flowering time	GWAS	Saidou et al.(2009)
Ryegrass	26	Diverse accessions	Heading date	GWAS	Skot et al.(2005)
Potato	221	Diverse accessions	Yield, quality	GWAS	Li et al.(2008)
	123	Diverse accessions	Disease resistance	CG	Malosetti et al.(2007)
Rice	577	Diverse landraces	Starch quality	CG	Bao et al.(2006)
	103	Diverse accessions	Yield	GWAS	Agrama et al.(2007)
	373	Inbred lines	14 agronomic traits	GWAS	Huang et al.(2010)
	118	Diverse accessions	Starch quality	CG	Yan et al.(2011)
Sorghum	195	Diverse accessions	Grain quality	CG	Figueiredo et al.(2010)
Soybean	139	Inbred lines	Iron deficiency	CG	Wang et al.(2008)
	96	Diverse accessions	Protein content	GWAS	Jun et al.(2008)
Sugar beet	111	Inbred lines	Yield, quality	GWAS	Stich et al.(2008b)
Wheat	95	Elite inbred lines	Kernel quality	GWAS	Breseghello et al. (2006)
	207	Elite inbred lines	Grain quality	GWAS	Reif et al.(2011)

[a] Nested association mapping (NAM) design

[b] Candidate gene-based or genome-wide association studies

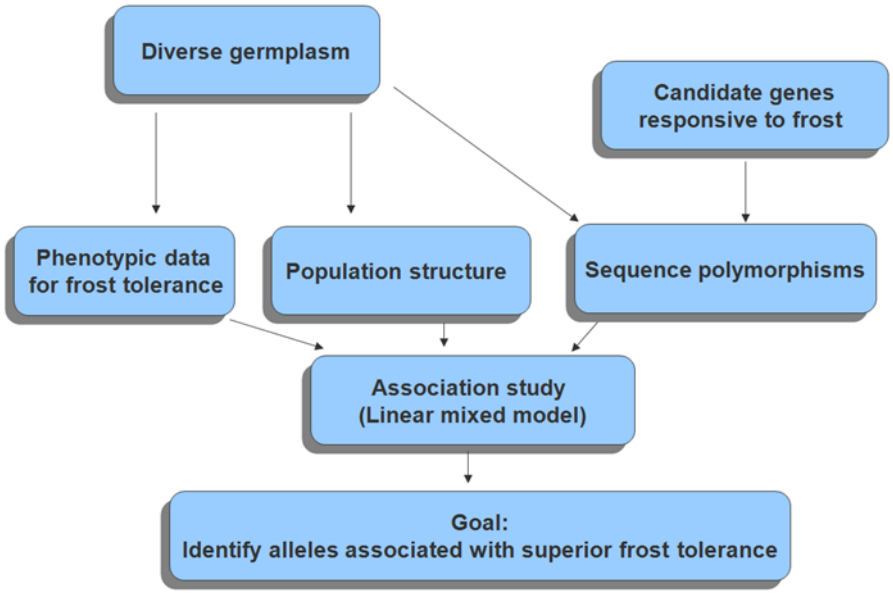


Figure 4 Schematic diagram of the candidate gene-based association analysis approach used in this study (modified after Zhu et al. 2008).

The basic principle of association studies is based on linkage disequilibrium (LD), also known as gametic phase disequilibrium, which is the non-random combination of alleles at different loci. LD is an established concept in theoretical and population genetics which is gaining more and more attention in the genomic era since it determines the marker density required for marker-based studies, such as association studies or genomic selection (Flint-Garcia et al. 2003). Many factors affect LD including linkage, selection, mutation, recombination, effective population size, and population structure (Flint-Garcia et al. 2003; Gupta et al. 2005). Studies on the extent of LD in various crops, such as *Triticum durum* (Maccaferri et al. 2005), *Zea mays* (Ching et al. 2002; Tenaillon et al. 2001), and *Sorghum bicolor* (Hamblin et al. 2005), indicated large variation in the extent of LD. The effect of germplasm on LD was clearly observed in barley, where LD decayed within 0.4 kb in wild material and extended up to 212 kb in elite lines (Caldwell et al. 2006a). LD decay can also vary considerably from locus to locus due to different recombination rates and selection pressures at different regions of the genome. In addition, higher levels of LD have been observed in self-pollinating species compared to outcrossing species, indicating that mating systems play a role (Flint-Garcia et al. 2003). Since rye is an outcrossing species, a low level of LD with a rapid decay is expected.

1.5 Objectives of this study

The major goal of this study was to identify alleles associated with superior frost tolerance (FT) in winter rye through a candidate gene-based association approach. Therefore, the objectives of this study were to

1) assess phenotypic variation of FT in five Middle and Eastern European winter rye populations using three different phenotyping platforms under controlled, semi-controlled and field conditions;
2) identify SNPs and haplotypes in twelve candidate genes with a putative important role in the frost responsive network;
3) examine genetic diversity and LD of these rye populations using genome-wide distributed SSRs and locus-specific SNPs from these twelve candidate genes;
4) investigate whether the SNPs and haplotypes of these twelve candidate genes are significantly associated with FT using linear mixed models;
5) investigate whether epistasis (gene by gene interactions) and SNP by environment interactions play a role in FT.

To address these objectives, an association panel of five Middle and Eastern European winter rye populations was established comprising 201 rye genotypes from important rye producing countries. Gamete capture was performed to allow for determining the linkage phase in the plants under study. Population structure and kinship were studied using 37 SSRs to avoid false positive results in the association studies.

2. Material and methods

2.1 Plant material

Plant material was derived from one Middle and four Eastern European cross-pollinated winter rye breeding populations. 33 plants from PR 2733 (Belarus), 44 plants from EKOAGRO (Poland), 15 plants from SMH2502 (Poland), 41 plants from ROM103 (Poland), and 68 plants from Petkus (Germany). For convenience, they will be hereafter referred to as PR, EKO, SMH, ROM, and Petkus, respectively. The Petkus population has undergone several cycles of recurrent selection, while the breeding history of the four Eastern European populations is unknown. Since rye is an outcrossing species, it is highly heterozygous, which leads to difficulties in determining haplotype phase. To address this problem, gamete capture was performed (Figure 5). Between 15 and 68 heterozygous plants from each of the five populations were crossed with the self-fertile inbred line Lo152 resulting in 201 heterozygous S_0 plants, each with one gamete known. To produce sufficient seed for phenotyping, S_0 plants were cloned in four to six plants per clone. S_0 plants were selfed to obtain S_1 families. 10-15 randomly chosen single plants from these S_1 families were subsequently selfed and the obtained seeds were bulked with equal amounts of seeds per single S_1 plant to obtain $S_{1:2}$ families. S_1 and $S_{1:2}$ families were used for phenotyping (see chapter 2.2). For molecular analyses, genomic DNA of S_0 plants was extracted from leaves according to the procedures in Rogowsky et al (1991).

2.2 Phenotypic data assessment

FT was measured in three phenotyping platforms: controlled, semi-controlled, and field. In the controlled platform, experiments were performed in climate chambers at -19°C and -21°C, in 2008 and 2009, respectively, at ARI Martonvásár, Hungary, using established protocols (Vagujfalvi et al. 2003). Briefly, seedlings were cold-acclimated in a six week hardening program with gradually decreasing temperatures from 15°C to -2°C. After that, plants were exposed to freezing temperatures within six days by decreasing the temperature from -2°C to -19°C or -21°C and then held at the lowest temperature for eight hours. After the freezing step, temperature was gradually increased to 17°C for regeneration. The ability of plants to re-grow was measured after two weeks using a recovery score, which ranged on a scale from 0: completely died, 1: little sign of life, 2: intensive damage, 3: moderate damage, 4: small damage, to 5: no damage. The experiment in 2008 contained 139 S_1 families. The experiment in 2009 contained 201 $S_{1:2}$ families, augmenting the same 139 S_1 families from the experiment in 2008 with an additional 62 $S_{1:2}$ families. Five plants of each S_1 or $S_{1:2}$ family

were grown as one test unit with five replicates per temperature and year. Due to the limited capacity of climate chambers, genotypes were randomly assigned into three and four chambers in 2008 and 2009, respectively.

In the semi-controlled platform, experiments in the two years 2008 and 2009 were performed with 3 replicates per year at Oberer Lindenhof, Germany, using the same 139 S_1 families and 201 $S_{1:2}$ families, respectively. From each family a test unit of 25 plants was grown outdoors in wooden boxes one meter above the ground in a randomized complete block design (RCBD). In case of snowfall, plants were protected from snow coverage to avoid damage by snow molds. Two weeks after a frost period, % leaf damage was scored as the proportion of the 25 plants of each family that showed leaf damage (dry and yellow leaves). In order to keep the same sign as with the measurements in the controlled and field platforms, % leaf damage was replaced by % plants with undamaged leaves, calculated as 100% - % leaf damage. Outcomes were recorded in January, February, and April of 2008 for the 139 S_1 families and in February and March of 2009 for the 201 $S_{1:2}$ families.

In the field platform, experiments were performed with the same 201 $S_{1:2}$ families in five environments in 2009 (Kasan, Russia, KAS; Lipezk, Russia, LIP1; Minsk, Belarus, MIN; Saskatoon, Canada, two different fields, SAS1 and SAS2) and in one environment in 2010 (Lipezk, Russia, LIP2). Depending on the environment 50-100 plants were used per test unit. The outcome, % survival, was calculated as the number of intact plants after winter divided by the total number of germinated plants before winter. RCBD design with 2 replicates was used for the SAS1 and SAS2 environments, while all other environments used 15 x 15 alpha-lattice design with 3 replicates.

2.3 Phenotypic data analyses

An analysis of variance (ANOVA) was performed separately for each environment from the three phenotyping platforms using the software package GenStat Discovery v3.0 (VSN International Ltd., Hemel Hempstead, United Kingdom). The correlations between environments based on raw means (means of replicates) of phenotypic values were calculated using Pearson's correlation.

To obtain adjusted entry means of each genotype, combined analyses of environments within each phenotyping platform were performed using linear mixed models. The general form for the three phenotyping platforms was:

$$y = 1\beta_1 + \boldsymbol{X_{PLATFORM}}\,\beta_{PLATFORM} + \boldsymbol{Z_{PLATFORM}}\,\gamma_{PLATFORM} + \boldsymbol{Z_{GENOTYPE}}\,\gamma_{GENOTYPE} + \varepsilon,$$

where y is the $n \times 1$ vector of platform-specific phenotypes, 1 denotes a $n \times 1$ vector of 1s and β_1 is the intercept. $\boldsymbol{X_{PLATFORM}}$ denotes the $n \times k$ design matrices for the fixed effects of platform, $\beta_{PLATFORM}$ is the associated fixed effects vector for the platform-specific effects, and $\boldsymbol{Z_{PLATFORM}}$ ($n \times m$), and $\boldsymbol{Z_{GENOTYPE}}$ ($n \times h$) are the corresponding design matrices for the random effects of platform (described in detail in 2.3.1, 2.3.2, and 2.3.3) and genotype, respectively. The residual error vector ε was assumed to comprise independent and identically distributed random Normal errors with mean of 0 and variance σ^2, $\varepsilon \sim N(0, \boldsymbol{I}\sigma^2)$. Statistical significance of fixed and random effects were assessed using Wald's test (Wald 1943) and the likelihood ratio test, respectively (Van Belle et al. 2004).

Broad-sense heritabilities (h^2) in the controlled and field platforms were estimated as:

$$\hat{h}^2 = \hat{\sigma}_g^{\,2} / (\hat{\sigma}_g^{\,2} + \hat{\sigma}_{ge}^{\,2}/t + \hat{\sigma}_e^{\,2}/rt)$$

where $\hat{\sigma}_g^{\,2}$, $\hat{\sigma}_{ge}^{\,2}$ and $\hat{\sigma}_e^{\,2}$ denote genotypic variance, genotype × environment interaction variance, and experimental error variance, respectively, and t and r are the number of environments and replications within environment, respectively. $\hat{h}^2$ for the semi-controlled platform could not be estimated using linear mixed model for longitudinal data due to unclear genotype × environment variance. The correlations between platforms based on adjusted entry means were calculated using Pearson's correlation coefficient.

2.3.1 Controlled platform

The outcome vector y was recovery score and the platform specific effect, $\beta_{PLATFORM}$ included the effect of year, temperature, genotype × year, and genotype × temperature interactions. A common platform-specific random effect controlling for the seven chambers across the two years 2008 and 2009 was included in the model, $\gamma_{PLATFORM} \sim N(0, \boldsymbol{I}\sigma^2_{chamber})$. No additional explicit generation adjustment for S_1 versus $S_{1:2}$ families was included in the statistical model as these were confounded with the fixed effect adjustment for year and the random chamber effects, and hence could not be additionally estimated. In other words, the generation effect was assumed implicitly adjusted for by other year effects in the model.

2.3.2 Semi-controlled platform

The outcome vector y was % plants with undamaged leaves measured repeatedly over three months (January, February, and April) in 2008 and two months (February, March) in 2009.

Therefore, a linear mixed model for longitudinal data was formulated where each test unit (25 plants) was assumed to be correlated among month of the same year. The platform specific effect vector, $\beta_{PLATFORM}$, included three terms: year effect, an overall linear trend in time for the three months in 2008 and two months in 2009, and the interaction of year and linear trend in time which would allow the linear time trend to differ in 2008 from 2009. Three platform specific random effects (vector $\gamma_{PLATFORM}$) were used: replication, a random intercept and a random slope with respect to month at each level of test unit. The random intercept and slope induce a correlation between two observations sharing the same random effect (test unit) and this correlation depends on time. That means in 2008 a test unit in January compared to the same test unit in February has a different (probably higher) correlation than this test unit in January compared to in April. Thus this model accounts for the correlated structure of repeated measurements.

2.3.3 Field platform

The outcome vector *y* was % survival and the platform-specific fixed effect $\beta_{PLATFORM}$ included indicator variables for the six environments, five environments in 2009 and one in 2010. Platform-specific random effects included genotype by environment interaction and a block effect nested within environments arising from the lattice design.

2.4 Candidate gene selection and primer design

Twelve candidate genes, *ScCbf2*, *ScCbf6*, *ScCbf9b*, *ScCbf11*, *ScCbf12*, *ScCbf14*, *ScCbf15*, *ScDhn1*, *ScDhn3*, *ScDreb2*, *ScIce2*, and *ScVrn1*, were selected based on their putative role on FT in closely related species. Individual *Cbf* genes were selected based on an expression study in rye (Campoli et al. 2009) and linkage mapping in barley and diploid wheat (Francia et al. 2007; Knox et al. 2008), *Vrn1* based on linkage mapping and a real-time PCR expression study in wheat (Galiba et al. 1995; Sutton et al. 2009), *Ice2* based on an expression study in wheat (Badawi et al. 2008), and *Dhn1* and *Dhn3* based on an expression study in barley (Choi et al. 1999). Following the *Cbf* nomenclature proposed by Skinner et al. (Skinner et al. 2005), names with the same number followed by different letters describe highly identical but distinct genes, for example, the highly identical *Cbf9a* and *Cbf9b* genes first identified by Jaglo et al. (2001). Primers for all genes were designed using Primer-BLAST from the NCBI database (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) based on sequences available in GenBank; information can be found in Table 2.

Table 2 Primer information on twelve candidate genes selected based on their association with frost tolerance

Primer set name	Forward (F) and reverse (R) primer sequence (5'-3')	PCR annealing temperature (°C)	Final $MgCl_2$ Concentration (mM)	Primer design based on (GenBank accession number)	GenBank accession number of submitted rye sequence
ScCbf2	F: CCTCGATCGGCCGGCGTGTAGC R: GTCCATGCCGCCGATCCAGTGCTC	66	1.5	*T. monococcum* (AY951945)	HQ730763
ScCbf6	F: ATGTGTCCGATCAAGAGGGA R: CTAGCTCTGGTAGCTCCAGA	60	1.5	*S. cereale* (EU194242)	HQ730764
ScCbf9b-fragment 1	F: TCTAGTGGTTGACGTGTGGG R: CGTCTCGTGGAACTTGGTC	62	1.5	*T. monococcum* (AY951945)	HQ730765
ScCbf9b-fragment 2	F: ACCACTACTCCACACCTCTCACGA R: TCCCCCAAAAGTAGAAACC	56	1.5	*S. cereale* (AF370730)	HQ730765
ScCbf11	F: ATGGAGTGGGCGTACAGCGG R: GTCAGTAGTTCCACAGGCTGA	63	1.5	*S. cereale* (EU194240)	HQ730766
ScCbf12-fragment 1	F: GCCTCAACTTCCCGGACT R: TCTTTCTTGTTTGCCAGCCT	52	1.5	*S. cereale* (EF028763)	HQ730767
ScCbf12-fragment 2	F: GCGTCCCGCAAAACTATAAA R: ATGTCGTGGCACAATGAGTC	63	2.0	*S. cereale* (EF028763)	HQ730767
ScCbf14	F: GTGATGGGCACAGGACG R: TTTCACAATGAACGAGCACG	65	1.5	*T. monococcum* (AY951945)	HQ730768
ScCbf15	F: AGCTCTCCTTCCTCTCCGTC R: GCCTTCAGTGTCCCAGCAC	64	1.5	*S. cereale* (EF028765)	HQ730769
ScDreb2-fragment 1	F: TGGAGCAGAGGAAAGTACCCGGA R: AGGTGGCTTCCTCGCCCTCT	65	1.5	*A. tauschii* (GU017675)	HQ730774
ScDreb2-fragment 2	F: CCAGCCTGGAAGGTGAGATCTTCTGT R: ATAGATGCCACTGGCGGCGCA	58	1.5	*A. tauschii* (GU017675)	HQ730774
ScDhn1	F: CCACGTAGCACGCACGCTGT R: TCTTCCTCCTCCCGCCCACG	61	1.5	*H. vulgare* (AF043087)	HQ730770
ScDhn3	F: TGGTGGGCATTTCCAGCCCG R: ACGTCCCGGGTACATACAAGCA	61	1.5	*H. vulgare* (AF043089)	HQ730771
ScIce2- fragment 1	F: GCACTTGATGGTGAATTTTGG R: TGATTGCGAACAAAAGCAAG	62	1.5	*T. aestivum* (EU562184)	HQ730772
ScIce2- fragment 2	F: TCCCTTCTCAGCTTGTTGAA R: GAGGAAGCTATTGGCTGTCG	62	1.5	*H. vulgare* (DQ113909)	HQ730772
ScVrn1	F: GGAGATTCGCACGTACGAT R: ATGACTCGGTGGAGAACTCG	58	1.5	*H. vulgare* (EU331765)	HQ730773

Due to limited information on rye DNA sequences in GenBank, primers for *ScVrn1*, *ScIce2*, *ScDhn1* and *ScDhn3* were designed based on homologous genes in *H. vulgare*, *T. aestivum*, and *T. monococcum*. Despite lack of homology in non-coding regions, putative functional regions of the candidate genes could be amplified. A 250 bp fragment of the promoter and first exon of *ScVrn1* was amplified since there is evidence that this region is one of the determinants of winter/spring growth habit in barley and wheat (Beales et al. 2005; Yan et al. 2003).

2.5 Amplification of candidate genes and DNA sequencing

Sixteen fragments of twelve candidate genes were amplified by PCR in 10 µl reaction volumes containing 10 ng DNA, 150 nM of each primer, 1x *Taq* DNA polymerase reaction buffer, 1.5 or 2.0 mM $MgCl_2$, 0.2 mM of each dNTP, and 0.5 U *Taq* DNA polymerase. After an initial denaturation at 96°C for 10 min, 35 cycles were conducted at 96°C for 1 min, primer-specific annealing temperatures at 52-66°C for 1 min, 72°C for 1 min, and a final extension step at 72°C for 15 min. The actual numbers of successful PCR amplification of the 201 genotypes differed from gene to gene ranging from 128 genotypes (64%) in *ScCbf11* to 198 (98%) in *ScVrn1*. Missing amplification products in individual genotypes were most likely the result of SNPs or insertion and deletion events (Indels) in the primer binding sites. However, absence of some *Cbf* genes in particular genotypes, as has recently been reported in barley and wheat (Fricano et al. 2009; Knox et al. 2010) cannot be excluded as an alternative explanation. The PCR products were purified in 96-well MultiScreen PCR plates (Millipore Corporation, Billerica, MA, USA) and directly sequenced through the QIAGEN sequencing service using Sanger sequencing technology (QIAGEN, Hilden, Germany) which has high sequencing accuracy of up to 99.99% (Shendure and Ji 2008). Amplicons of each S_0 plant were sequenced with both forward and reverse PCR primers.

2.6 Sequence analyses

Sequence data were assembled into contigs and SNPs were detected using the software Variant Reporter™ V1.0 (Applied Biosystems, Foster City, CA, USA). The DNA sequence of the homozygous inbred line Lo152 was used as reference sequence, and alleles of this common parent were subtracted from all sequences to determine the haplotype phase. Heterozygous Indels were detected either manually by checking sequences from both strands or using the web-based program Indelligent v1.2 (Dmitriev and Rakitov 2008). In case of large Indels, amplicons from the respective genotypes were sub-cloned using the TOPO TA

Cloning Kit (Invitrogen, Carlsbad, CA, USA). At least five clones were sequenced to resolve large heterozygous Indels. Sequences of the Lo152 reference alleles from the twelve candidate genes were submitted to GenBank under accession numbers HQ730763–HQ730774. For convenience, polymorphic sites along the sequence were numbered starting with "SNP1". Indels were treated as single polymorphic sites and referred to in the text as SNPs instead of differentiating between SNPs and Indels. For all sequence analyse, Lo152 alleles were excluded leading to a situation similar to inbred line. Haplotypes and haplotype frequencies were determined within each candidate gene using Arlequin v3.1 (Excoffier et al. 2005). Nucleotide diversity (π) was calculated as the average number of nucleotide differences per site between two sequences for both, the complete sequences and restricted to exons. Haplotype diversity (*Hd*) was calculated as the probability that two randomly chosen haplotypes from a given population were different. Analyses of nucleotide and haplotype diversity were performed separately for each population as well as for all populations grouped together using the software DnaSP v5.10 (Rozas et al. 2003). Average nucleotide diversity (π) over all genes was calculated using concatenated sequences that were generated in software TASSEL v2.1(Bradbury et al. 2007). To test for selection Tajima's *D* was calculated as the difference between the mean pairwise nucleotide differences (π) and the number of segregating sites (*S*) relative to their standard error using the software DnaSP v5.10. The statistical significance of Tajima's *D* was obtained assuming that *D* follows a beta distribution.

2.7 SSR genotyping and genetic diversity analyses

Thirty seven SSR markers were chosen from literature based on their experimental quality and map location as providing even coverage of the rye genome. Primers and PCR conditions for rye microsatellite (RMS) and *Secale cereale* microsatellite (SCM) markers were described in detail by Khlestkina et al. (2004) and Hackauf and Wehling (2002), respectively. Fragments were separated using a 3130xl Genetic Analyzer (Applied Biosystems Inc., Foster City, CA, USA) and allele sizes were assigned using the program GENEMAPPER (Applied Biosystems Inc., Foster City, CA, USA). Genotyping data obtained from the SSR analyses of the 201 genotypes were used for the following calculations. Polymorphic information content (PIC) was estimated according to Botstein et al. (1980) using the software package PowerMarker v3.0 (Liu and Muse 2005). PIC, the probability that a given offspring of a parent carrying the rare allele at the index locus will allow deduction of the parental genotype at the marker locus, was estimated as

$$PIC_l = 1 - \sum_{u=1}^{g} p_{lu}^2 - \sum_{u=1}^{g-1} \sum_{v=u+1}^{g} 2\, p_{lu}^2\, p_{lv}^2 ,$$

where g is number of alleles, p_{lu} and p_{lv} are the frequencies of the uth and vth alleles at the lth locus, respectively. To eliminate bias whereby the observed number of alleles highly depends on the number of analysed genotypes, allelic richness (R_s) was estimated from a rarefaction method (Petit et al. 1998) implemented in FSTATv2.9.3 (Goudet 1995). Briefly, the method estimates the expected number of alleles in a sub-sample of n genotypes, given that N genotypes have been sampled at a locus, where N ≥ n. In this study it was specifically calculated as

$$R_s = \sum_{i=1}^{s} \left[1 - \frac{\binom{N - N_i}{n}}{\binom{N}{n}} \right]$$

where N is the number of observed genotypes (201 or less), N_i the number of genotypes with type i alleles among the N genotypes, n the number of genotypes in each population, and S is the total number of alleles among the N genotypes. To visualize the degree of variation within and between populations, principal co-ordinate analysis (PCoA) was performed based on SSRs and haplotypes of candidate genes using DICE similarity coefficients (Dice 1945) estimated in NTSYSpc v2.2 (Applied Biostatistics Inc., Setauket, NY, USA). Analysis of molecular variance (AMOVA) was performed based on SSRs and candidate gene polymorphisms using Arlequin v3.1 with 15,000 permutations of the data to estimate statistical significance at $P < 0.001$ for each variance component (Excoffier et al. 2005; Excoffier et al. 1992). The Lo152 alleles were excluded from all analyses.

2.8 Estimation of population structure and kinship

In order to correct for confounding effects in the association studies, population structure and kinship were estimated based on 37 SSR markers as described in 2.7. Population structure was inferred using the STRUCTURE software v2.2, which is based on a Bayesian model-based clustering algorithm that incorporates admixture and allele correlation models to account for genetic material exchange in populations resulting in shared ancestry (Pritchard et al. 2000a). Briefly, the method assigned each individual to a predetermined number of groups

(*k*) characterized by a set of allele frequencies at each locus, assuming that the loci are in Hardy-Weinberg equilibrium and linkage equilibrium. Ten runs for values of *k* ranging from two to eleven were performed using a burn-in period of 50,000 replications followed by 50,000 Markov Chain Monte Carlo iterations. Posterior probabilities of each *k* were averaged over the ten runs to determine the maximum posteriori *k*. The population structure matrix $\boldsymbol{Q}_{\boldsymbol{STRUCTURE}}$ was estimated, providing for each of the 201 genotypes an estimate of the proportion membership in the *k* populations. The phylogenetic tree was constructed based on Rogers distance (Rogers 1972) from 37 SSR markers using a clustering method UPGMA (Unweighted Pair Group Method with Arithmetic Mean) implemented in PowerMarker v3.25 (Liu and Muse 2005). The kinship matrix (***K***) was estimated from the same SSR markers using the allele-similarity method (Hayes and Goddard 2008), which guarantees a positive semi-definite relationship matrix among the 201 genotypes, and was used for the covariance structure of the random genotype effects in the linear mixed model. For a given locus, the similarity index S_{xy} between two genotypes was 1 when alleles were identical and 0 when alleles were different. S_{xy} was averaged over the 37 loci and transformed and standardized as $\hat{S}_{xy} = (S_{xy} - \hat{S}_{xymin}) / (1 - \hat{S}_{xymin})$, where $\hat{S}_{xymin}$ is the minimum relationship in the matrix.

2.9 Linkage disequilibrium

Linkage disequilibrium was measured by the parameter r^2 (Hill and Robertson 1968) for candidate genes and SSR markers using DnaSP v5.10 and TASSEL v2.1, respectively, with Indels treated as single polymorphic sites and SNPs with minor allele frequencies (MAF) < 0.05 excluded due to instability. r^2 is calculated as

$$r^2 = \frac{(p_{AB}p_{ab} - p_{Ab}p_{aB})^2}{p_A p_B p_a p_b},$$

where p_A, p_B, p_a, and p_b are the frequencies of alleles A, B, a, and b in the population. Haplotype frequencies of allele combinations are denoted as p_{AB}, p_{Ab}, p_{aB}, and p_{ab}, respectively. The LD decay curve was fitted under the mutation-drift-equilibrium model, the expected value of r^2 is E (r^2) = 1 / (1+4*Nc*), where *N* is the effective population size, and *c* is the recombination fraction between sites. With assumption of a low mutation rate and an adjustment for sample size, the expectation becomes (Hill and Weir 1988):

$$E(r^2) = \left[\frac{10+\Gamma}{(2+\Gamma)(11+\Gamma)}\right]\left[1+\frac{(3+\Gamma)(12+12\Gamma+\Gamma^2)}{n(2+\Gamma)(11+\Gamma)}\right],$$

where $\Gamma=4Nc$ and n is the number of genotypes compared. The LD decay curve was estimated using a non-linear least-squares estimate of Γ fitted by the *nls* function in the R software package (R Development Core Team 2009), separately for each population and for all populations pooled together. The approach of Breseghello and Sorrells (2006) was used to determine threshold values of r^2 that indicated significant LD. r^2 values were estimated from 37 unlinked SSR markers and square root transformed so that they would be better approximated by a Normal distribution. The 95th percentile from the empirical distribution of all 666 pairwise r^2 derived from the 37 genome-wide distributed SSR markers was selected as the threshold value, with the rationale that any value above the threshold could in high likelihood be attributable to genetic linkage. Threshold values were calculated separately for each population and for all populations pooled together. The extent of LD of SNPs within candidate genes was estimated as the point where the LD decay curve passed below the threshold.

2.10 SNP-FT association analyses

Twelve candidate genes *ScCbf2*, *ScCbf6*, *ScCbf9b*, *ScCbf11*, *ScCbf12*, *ScCbf14*, *ScCbf15*, *ScDhn1*, *ScDhn3*, *ScDreb2*, *ScIce2*, and *ScVrn1* were tested for association with FT. SNP-FT associations in all platforms were performed using linear mixed models that evaluated the effects of SNPs with MAF > 5% individually, adjusting for population structure, kinship and platform-specific effects. A one stage approach was chosen for analysis which directly models the phenotypic raw data as the response variable. The models for each phenotyping platform were the same as in phenotypic data analyses plus the factors of SNPs, population structure, and kinship. The general form of the linear mixed model for the three platforms was:

$$y = 1\beta_1 + \boldsymbol{X_{SNP}}\beta_{SNP} + \boldsymbol{Q_{STRUCTURE}}\,\beta_{STRUCTURE} + \boldsymbol{X_{PLATFORM}}\,\beta_{PLATFORM} + \boldsymbol{Z_{PLATFORM}}\,\gamma_{PLATFORM} + \boldsymbol{Z_{GENOTYPE}}\,\gamma_{GENOTYPE} + \varepsilon,$$

where y is the $n \times 1$ vector of platform-specific phenotypes, 1 denotes a $n \times 1$ vector of 1s and β_1 is the intercept, $\boldsymbol{X_{SNP}}$ $(n \times p)$, $\boldsymbol{Q_{STRUCTURE}}$ $(n \times q)$ and $\boldsymbol{X_{PLATFORM}}$ $(n \times k)$ are design matrices for the fixed effects of SNPs, population membership and platform, respectively, and $\boldsymbol{Z_{PLATFORM}}$ $(n \times m)$ and $\boldsymbol{Z_{GENOTYPE}}$ $(n \times h)$ are the corresponding design matrices for the random effects of platform (described in 2.3.1, 2.3.2, and 2.3.3) and genotype. If a platform contained random effects, these were accommodated by including a random effect $\gamma_{PLATFORM}\sim$

N (0, $\boldsymbol{D}\sigma^2$) with mean of 0, and variance covariance matrix $\boldsymbol{D}$. The random genotype effect was similarly assumed to follow a Normal distribution, $\gamma_{GENOTYPE} \sim$ N (0, $2\boldsymbol{K}\sigma^2_g$), where $\boldsymbol{K}$ was the estimated kinship matrix and σ^2_g the variance component due to genotype. In order to account for kinship in the estimation of random genotype effects, $\gamma_{GENOTYPE}$, the design matrix $\boldsymbol{Z_{GENOTYPE}}$ was multiplied by the cholesky-root of the kinship matrix. The residual error vector ε ($n \times 1$) was assumed to comprise independent and identically distributed random Normal errors with mean of 0 and variance σ^2, $\varepsilon \sim$ N (0, $\boldsymbol{I}\sigma^2$).

Analyses of SNP-FT associations were performed using the lme4 package (Bates and Maechler 2010) implemented in R (R Development Core Team 2009). Significance of individual SNP effects was assessed via the *t*-statistic performed at the two-sided alpha = 0.05 level. A multiple testing problem arises, which inflates the false positive rate of the study. A simple and common way to handle this problem is Bonferroni correction where the significance level is divided by the number of tests. However, the Bonferroni correction is too conservative and only suitable for independent tests, an assumption violated in this study due to a high LD between some of the SNPs as shown in chapter 3.4. Therefore, the less stringent significance level of alpha = 0.05 is reported in the main results in order to retain candidates for further validation in upcoming experiments. Empirical correlations between the 170 SNP-FT associations reported among the three phenotyping platforms were performed using Pearson's correlation based on the *t* values from the corresponding association tests. The genetic variation explained by an individual SNP or haplotype was calculated as $100 \times ((\sigma^2_g - \sigma^2_{gSNP})/\sigma^2_g)$, where σ^2_g is the genetic variation in the reduced model without an individual SNP and σ^2_{gSNP} is the model including an individual SNP (Mathews et al. 2008). This ad-hoc measure can result in negative estimates since variance components do not automatically decrease with more adjustment in a model as error sums of squares do; negative estimates were truncated to 0. Genetic variation explained by non-significant SNPs was also reported since these 170 SNPs were chosen from candidate genes with prior knowledge. SNP × environment interaction effects were tested by likelihood ratio test, comparing the full model with main effects of SNP and environment plus interaction between them to the reduced model with main effects of SNP and environment only. The null distribution of this likelihood ratio test is asymptotic to the chi-square distribution (Van Belle et al. 2004).

2.11 Haplotype-FT association and epistasis

Haplotype phase was determined by subtracting the common parent Lo152 alleles and haplotypes were defined within each candidate gene using DnaSP v5.10 (Rozas et al. 2003).

Haplotype-FT associations were performed using candidate gene haplotypes with MAF > 5%. The same platform-specific statistical models controlling for population structure, kinship, and platform-specific effects were used to test associations between haplotypes of the respective candidate genes and FT. For these analyses β_{hap} replaced β_{SNP} as a measure of the haplotype effect of the non-reference compared to the reference haplotype Lo152. First, significant differences between haplotypes of one gene were assessed using the ANOVA F-test. Under the null hypothesis of no association between haplotypes and FT, the effects β_{hap} for each haplotype will be the same. If the overall statistic was significant, individual haplotype effects were tested against the reference haplotype Lo152 via *t*-tests. Based on haplotype information gene × gene interactions (with frequency of haplotype combination > 5%) were assessed using the likelihood ratio test, comparing the full model with main effects plus interaction to the reduced model with main effects only. The null distribution of this likelihood ratio test is asymptotic to the chi-square distribution (Van Belle et al. 2004).

3. Results

3.1 Phenotypic variation in the three phenotyping platforms

Phenotypic assessment of FT was carried out in four environments in the controlled, two environments in the semi-controlled, and six environments in the field platforms. Phenotypic data were analyzed separately in each environment which revealed significant genotypic variation in ten out of twelve environments from the three different phenotypic platforms (Table 3).

Table 3 Means over replicates of recovery score, % plants with undamaged leaves, and % survival, variance components, and F test for each environment in the controlled, semi-controlled, and field platform, respectively.

Environment	Source	Mean	Variance components	*P*-value from F test
Controlled platform				
-19°C/2008	Genotype	2.42	0.233	<0.01
	Chamber		0.004	<0.01
	Residual		0.449	
-21°C/2008	Genotype	1.94	0.177	<0.01
	Chamber		0.934	<0.01
	Residual		0.430	
-19°C/2009	Genotype	1.53	0.147	<0.01
	Chamber		0.060	<0.01
	Residual		0.507	
-21°C/2009	Genotype	1.09	0.168	<0.01
	Chamber		0.048	<0.01
	Residual		0.514	
Semi-controlled platform				
January 2008	Genotype	92.59	8.10	<0.01
	Residual		15.77	
February 2008	Genotype	75.08	15.94	<0.01
	Residual		45.84	
April 2008	Genotype	47.04	25.90	<0.01
	Residual		75.64	
February 2009	Genotype	67.14	15.67	<0.01
	Residual		85.42	
March 2009	Genotype	35.52	9.74	<0.01
	Residual		30.31	
Field platform				
KAS	Genotype	62.4	2.96	0.52
	Block		10.85	0.03
	Residual		34.46	
LIP1	Genotype	65.7	6.37	<0.01
	Block		35.33	<0.01
	Residual		66.87	

(Table 3 continued)

Environment	Source	Mean	Variance components	*P*-value from F test
MIN	Genotype	74.1	0.00	0.90
	Block		5.27	<0.01
	Residual		37.60	
SAS1	Genotype	69.4	400.40	<0.001
	Residual		531.64	
SAS2	Genotype	26.0	248.56	<0.001
	Residual		545.65	
LIP2	Genotype	85.1	14.03	<0.01
	Block		17.33	<0.01
	Residual		71.93	

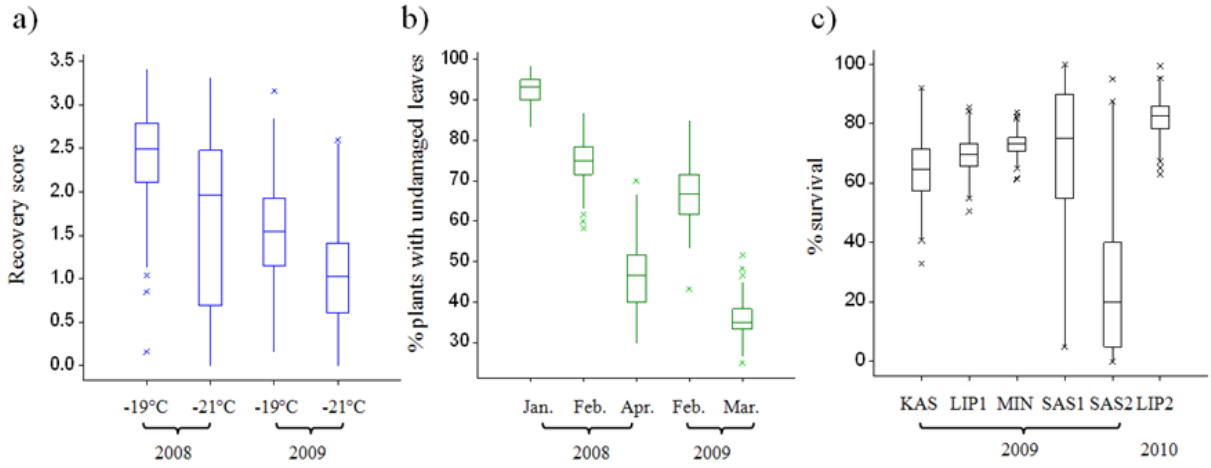


Figure 5 Phenotypic variation in a) controlled, b) semi-controlled, and c) field phenotyping platforms in the years 2008, 2009, and 2010.

The values are the average phenotypic raw values of three replicates (two for SAS1 and SAS2) for each genotype. Boxes indicate the range of the middle 50% of the data with a horizontal line representing the median and vertical lines beyond the boxes indicate the upper and lower 25% of the phenotypic means. Outliers are represented by crosses. -19°C and -21°C are the two tested temperatures. Jan., Feb., Apr., and Mar. are months when FT was measured. KAS, LIP1, MIN, SAS1, SAS2, and LIP2 are the field locations.

In the controlled platform, genotypic variation for FT was significant at both temperatures for both years ($P < 0.01$). The median recovery scores ranged from 2.5 (between intensive and moderate damage) at -19°C in 2008 to 1.0 (little sign of life) at -21°C in 2009 (Figure 5). As expected, recovery scores were higher at -19°C than at -21°C in the same year but were lower in 2009 than in 2008 probably due to different generations of rye material and thus different levels of inbreeding (S_1 vs $S_{1:2}$ families). The high variability at -21°C in 2008 might have been induced by substantial variation between chambers. The correlations of FT between -19°C and -21°C in 2008 was 0.59 which was similar to that in 2009 ($r = 0.60$).

In the semi-controlled platform, genotypic variation for FT was significant during all months for both years ($P < 0.01$). Linear decreasing trends were observed during each year which was expected since those were longitudinal data and thus the damaged portions of plants increased during the progression of winter. The correlations of FT between each month in the same year were low to medium ranging from $r = 0.22$ (January 2008 vs April 2008) to $r = 0.51$ (January 2008 vs February 2008).

In the field platform, genotypic variation for FT was significant ($P < 0.01$) in four (LIP1, LIP2, SAS1, and SAS2) of the six environments. Compared to other environments, SAS1 and SAS2 showed much larger genotypic variances (Table 3) and better differentiations for FT among genotypes, with a survival rate ranging from 5% to 100% with a median of 75% and 0% to 95% with a median of 20%, respectively. The large difference of survival rate between SAS1 and SAS2 was probably due to different altitudes and consequently different severity of frost stress. The correlations of FT between each environment were very low ranging from $r = 0.01$ (KAS vs SAS1) to $r = 0.36$ (SAS1 vs SAS2).

In combined analyses across environments in each phenotyping platform, the adjusted entry means for recovery score of the 201 genotypes in the controlled platform ranged from 1.02 to 2.24, with a mean of 1.73 indicating scores between little sign of life and intensive damage (Table 4). Significant differences ($P < 0.01$) between genotypes, temperatures, and chambers were observed. In the semi-controlled platform, the adjusted entry means of % plants with undamaged leaves in the 201 genotypes ranged from 56.0% to 70.9%, with a mean of 63.5%. The effects of genotype and year were significant ($P < 0.01$). In the field platform, the adjusted entry means of % survival in the 201 genotypes varied between 57.4% and 69.4%, with an average of 62.7%. The effects of genotype, environment and genotype × environment interaction were significant ($P < 0.01$). The heritability estimates were low in the field platform ($\hat{h}^2 = 0.35$), but higher in the controlled platform ($\hat{h}^2 = 0.67$). As no estimation of variance component for genotype x environment was possible for longitudinal data due to confounded genotype × environment variance, calculation of $\hat{h}^2$ for the semi-controlled platform was not performed.

Table 4 Adjusted entry means (minimum-maximum), variance components, and heritabilities ($\hat{h}^2$) of FT in three phenotyping platforms in the combined analyses across environments

Source	Phenotypic means [a]	Variance components	*P*-value [b]	$\hat{h}^2$
Controlled				
Genotype	1.73 (1.02-2.24)	0.14	<0.001	0.67
Temperature		NA	<0.001	
Year		NA	<0.001	
Genotype × temperature		0.09	<0.01	
Genotype × year		0.06	<0.01	
Chamber		0.10	<0.001	
Residual		0.61		
Semi-controlled				
Genotype	63.5 (56.0-70.9)	7.70	<0.001	NA [c]
Year		NA	<0.001	
Overall linear trend		NA	<0.001	
Year × overall linear trend		NA	<0.001	
Linear trend per test unit		37.1	<0.001	
Residual		46.0		
Field				
Genotype	62.7 (57.4-69.4)	14.4	<0.001	0.35
Environment		NA	<0.001	
Genotype × environment		68.9	<0.001	
Block		10.9	<0.01	
Residual		239.1		

NA: not available
[a] Recovery score (0-5) in the controlled platform; % plants with undamaged leaves in the semi-controlled platform; % survival in the field platform
[b] Wald's test for fixed effects and likelihood ratio test (LRT) for random effects (see chapter 2.3)
[c] Not estimated due to confounded genotype × environment variance

When looking at FT in each population, the PR population showed the highest mean FT over the three platforms (Table 5). Among the top ten most frost tolerant genotypes, six were from the PR population; among the top ten most frost susceptible genotypes, five were from the Petkus population. However, there was a significant differentiation of FT among populations in the three platforms ($P < 0.01$). The correlation between adjusted entry means of FT was low with $r = 0.38$ ($P < 0.01$) between the controlled and semi-controlled platforms; $r = 0.31$ ($P < 0.01$) between the controlled and field platforms; and $r = 0.19$ ($P < 0.01$) between the semi- controlled and field platforms (Figure 6).

Table 5 Adjusted entry means ± SE and range (minimum-maximum) for recovery score, % plants with undamaged leaves and % survival in the controlled, semi-controlled and field platform in each population and over all populations, respectively.

	Overall	PR	EKO	SMH	ROM	Petkus
No. of genotypes	201	33	44	15	41	68
Controlled	1.73±0.01 (1.02-2.24)	1.82±0.04 (1.30-2.24)	1.78±0.02 (1.40-2.02)	1.66±0.06 (1.33-2.15)	1.69±0.03 (1.02-2.05)	1.70±0.02 (1.34-2.07)
Semi-controlled	63.5±0.18 (56.0-70.9)	64.6±0.36 (60.1-68.5)	64.3±0.37 (59.0-70.9)	63.3±0.75 (56.0-67.6)	64.4±0.32 (58.3-67.7)	62.1±0.26 (57.6-66.4)
Field	62.7±0.17 (57.4-69.4)	64.1±0.44 (57.5-69.4)	63.1±0.31 (57.8-67.8)	61.1±0.62 (58.1-64.5)	62.0±0.35 (58.1-67.1)	62.4±0.27 (57.4-67.1)

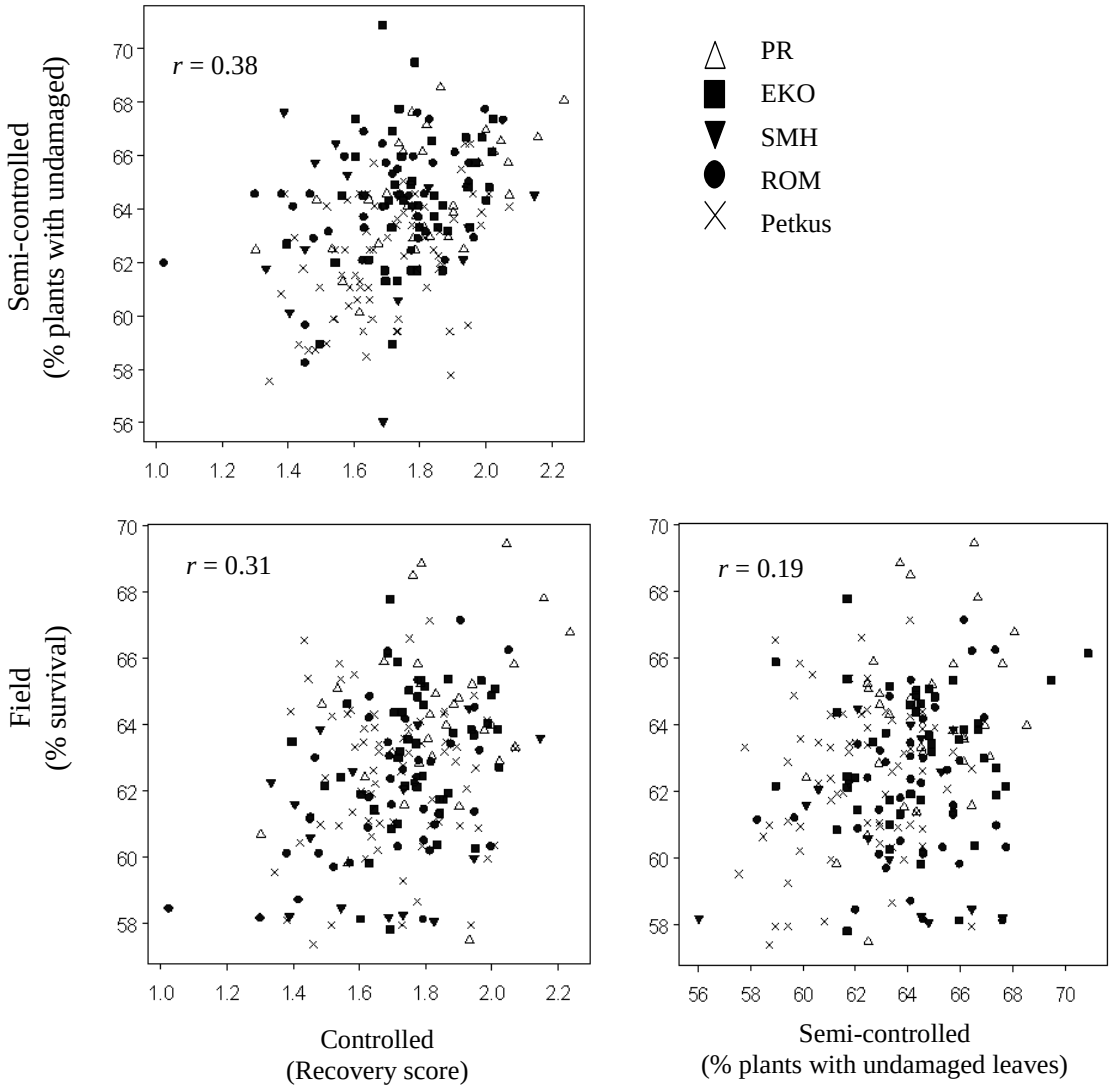


Figure 6 Scatter plots showing correlations between the three phenotyping platforms based on adjusted entry means of 201 genotypes.

3.2 Genetic diversity

3.2.1 Assessing genetic diversity of the five rye populations using SSR markers

Genetic diversity within the five populations was summarised based on 37 genome-wide SSR markers (Table 6). A total of 230 alleles and an average of 6.2 alleles per locus were observed. PIC varied from 0.37 to 0.51 with an average of 0.47. Allelic richness, which is not affected by sample size, ranged from 2.51 to 3.43, with a mean of 3.16. PIC was highly correlated with allelic richness (r = 0.965). Compared to the four Eastern European populations, the Petkus population had a slightly lower mean number of alleles per locus, PIC, allelic richness, and number of private alleles, despite the fact that it had the largest population size. Genetic diversities of individual SSR markers across the five populations are summarized in Table 7. The 37 SSR markers were evenly distributed across the whole rye genome with on average five SSR markers per chromosome with an average marker interval of 21 cM according to the length of the integrated consensus map of Gustafson et al. (2009). The number of alleles of each SSR marker was high ranging from 2 to 23. PIC among 37 SSR markers varied from 0.12 to 0.85 with an average of 0.47 while allelic richness from 1.67 to 6.68 with an average of 3.42. The top five highly polymorphic SSR markers are RMS1007, RMS1121, RMS1083, RMS1012, and SCM214 in terms of allelic richness.

Table 6 Genetic diversity within populations based on 37 SSR markers

Population	No. of genotypes	No. of private alleles [a] (%)	Average no. of alleles (range)	PIC [b]	Allelic richness [c]
PR	33	20 (12.1%)	4.46 (2-12)	0.50	3.43
EKO	44	14 (8.8%)	4.30 (2-18)	0.49	3.28
SMH	15	3 (2.4%)	3.38 (1-9)	0.46	3.18
ROM	41	13 (7.7%)	4.50 (2-13)	0.51	3.38
Petkus	68	4 (3.6%)	3.00 (1-10)	0.37	2.51

[a] Private alleles denotes the number of alleles which occurred only in one population

[b] PIC: Polymorphic information content, a higher value means higher genetic diversity

[c] Allelic richness is a measure of the number of alleles independent of sample size, a higher value means higher genetic diversity

Table 7 Chromosomal locations and genetic diversities estimated as numbers of alleles, polymorphic information content (PIC) and allelic richness of the 37 SSRs

SSR markers	Chromosome	No. of alleles	PIC	Allelic richness
RMS1107	1R	7	0.59	3.64
RMS1280	1R	4	0.16	2.18
RMS1303	1R	5	0.29	2.34
SCM247	1R	4	0.42	2.61
SCM266	1R	2	0.16	2.04
RMS1138	2R	7	0.35	3.59
RMS1230	2R	5	0.69	4.08
RMS1238	2R	5	0.60	3.52
SCM276	2R	3	0.27	2.58
SCM290	2R	5	0.41	3.23
SCM294	3R	3	0.33	2.14
RMS1028	3R	9	0.55	4.26
RMS1254	3R	6	0.55	3.55
RMS1261	3R	10	0.47	3.77
RMS1323	3R	4	0.08	1.67
RMS1007	4R	23	0.85	6.68
RMS1026	4R	5	0.19	2.16
RMS1181	4R	2	0.32	2.00
SCM047	4R	2	0.19	1.94
RMS1083	5R	16	0.80	5.98
RMS1205	5R	5	0.48	2.81
RMS1218	5R	3	0.57	3.42
RMS1237	5R	8	0.56	3.72
RMS1259	5R	7	0.72	4.49
RMS1278	5R	4	0.53	2.96
SCM260	5R	5	0.13	2.20
RMS1090	6R	6	0.57	3.24
RMS1121	6R	15	0.63	6.44
SCM107	6R	2	0.22	1.97
SCM214	6R	8	0.79	5.22
RMS1012	7R	14	0.78	5.79
RMS1018	7R	11	0.77	5.34
RMS1187	7R	2	0.36	2.00
RMS1188	7R	4	0.52	3.17
RMS1197	7R	4	0.36	2.82
SCM063	7R	3	0.44	2.93
SCM322	7R	3	0.58	4.06
Mean		6.2	0.47	3.42

3.2.2 Sequence polymorphisms and genetic diversity in candidate genes

In total, 9,316 bp from 12 candidate genes in 201 rye genotypes were amplified resulting in 161 SNPs, nine Indels, and an average polymorphism frequency of 1 polymorphism / 55 bp (Table 8). Forty-three SNPs were non-synonymous polymorphisms resulting in amino acid

replacements, 17 of which changed polarity. In the *Cbf* gene family, *ScCbf9b* had the highest number of SNPs (N=30), of which ten were non-synonymous and three led to an exchange of amino acids of different polarity. The first intron and second exon comprising 20% of the coding sequence of *ScIce2* were amplified, resulting in the identification of 37 SNPs, all located in the first intron. A 250 bp fragment of the promoter and first exon of *ScVrn1* was amplified but no polymorphic site was identified, except for a 2 bp Indel. Out of nine Indels identified, seven were located in the non-coding regions of *ScCbf2*, *ScCbf9b*, *ScVrn1*, *ScDhn1*, and *ScDhn3* and two in the coding regions of *ScCbf12* and *ScCbf15* without causing a frame shift. It is noteworthy that the 200 bp Indel in the promoter of *ScCbf2* contained two MYB and one MYC *cis*-elements, putative binding sites for the *ScIce2* transcription factor.

Nucleotide diversity (π) ranged from 0.4×10^{-3} in *ScVrn1* to 14.5×10^{-3} in *ScCbf11*, and when restricted to exons, from 0 in *ScIce2* and *ScVrn1* to 14.5×10^{-3} in *ScCbf11* (Table 8). The biggest difference between analyses of π for the whole gene compared to restriction to exons occurred in *ScIce2* where π decreased from 11.2 to 0 due to absence of SNPs in the exon. Haplotype diversity (*Hd*) ranged from 0.11 in *ScVrn1* to 0.98 in *ScCbf9b*. A significant positive Tajima's *D* value was observed over all populations for *ScCbf15* and *ScIce2*, whereas a significant negative value was observed in *ScDhn1* and *ScDreb2*. In the SMH population, *ScCbf6*, *ScIce2*, and *ScDhn1* had reduced nucleotide and haplotype diversities. Similarly in the PR and EKO populations, respectively, *ScCbf11* and *ScCbf15* had reduced nucleotide and haplotype diversities compared to the other genes (Additional table 1). Haplotype frequencies varied markedly between candidate genes, with some candidate genes dominated by a single haplotype and others with a more balanced haplotype frequency distribution (Figure 7). For example, in *ScCbf14*, *ScVrn1*, and *ScDhn1*, the most frequent haplotype occurred in more than 70% of genotypes, whereas in *ScCbf9b* all haplotypes occurred with frequencies less than 10%. The finding in *ScCbf9b* can be attributed to a large number of haplotypes (N=95) with high haplotype diversity primarily generated by polymorphic sites located in the coding region. Similarly, only five of 48 haplotypes in *ScCbf12* occurred at a frequency greater than 10%. For *ScCbf14*, all populations had a similar distribution of haplotype frequencies. However, for *ScCbf15* haplotypes 1, 2, 3, and 4 were evenly distributed in PR, whereas in the other four populations only two haplotypes (EKO and SMH: 1 and 2; ROM and Petkus: 1 and 4) were prevalent (80% - 95%). For *ScCbf11*, haplotype 1 was predominant in the PR and Petkus populations, occurring in 82% and 57% of genotypes, respectively, whereas haplotype 2 predominated in EKO (67%) and SMH (75%).

Table 8 Summary information on candidate gene (CG) sequences and sequence analysis

Analyzed fragment length, gene coverage, number of genotypes, number of SNPs (MAF > 0.05), number of Indels and haplotypes (MAF > 0.05), haplotype (*Hd*) and nucleotide (π) diversity, Tajima's *D*, and linkage disequilibrium (LD)

CG	Fragment length (bp)	Gene coverage [a]	No. of genotypes [b]	No. of SNPs (non-synonymous)	No. of Indels	No. of haplotypes	*Hd*± SD	π± SD ×10^{-3} (only exon)	Tajima's *D* [d]	Average LD (r^2)
ScCbf2	619	5'UTR/E	169	2 (0)	1	7 (77.3%)[c]	0.67 ± 0.02	1.5 ± 0.1 (1.4 ± 0.1)	1.17	0.13
ScCbf6	495	E	197	3 (0)	0	9 (88.8%)	0.44 ± 0.04	3.6 ± 0.3	-0.35	0.77
ScCbf9b	1,371	5'UTR/E/3'UTR	183	30 (10)	1	95 (12.1%)	0.98 ± 0.03	7.1 ± 0.3 (11.5 ± 0.2)	1.71	0.14
ScCbf11	623	E	128	27 (12)	0	12 (60.0%)	0.65 ± 0.02	14.5 ± 0. 9	1.74	0.51
ScCbf12	754	5'UTR/E/3'UTR	141	25 (8)	1	48 (39.7%)	0.89 ± 0.02	8.8 ± 1.0 (7.7± 0.1)	0.40	0.38
ScCbf14	560	E	185	5 (3)	0	4 (89.7%)	0.17 ± 0.04	1.5 ± 0.3	-0.27	0.92
ScCbf15	502	E	172	3 (3)	1	9 (73.8%)	0.68 ± 0.04	3.0 ± 0.2	2.14*	0.30
ScDhn1	435	5'UTR/E	138	4 (1)	2	12 (53.4%)	0.33 ± 0.05	2.7 ± 0.5 (4.4± 0.1)	-1.86*	0.48
ScDhn3	514	I / E/3'UTR	130	12 (2)	2	21 (83.2%)	0.73 ± 0.03	8.1 ± 0.6 (8.9± 0.1)	0.008	0.25
ScDreb2	1,677	E3/I3/E4	197	13 (4)	0	47 (57.8%)	0.89 ± 0.02	3.0 ± 0.3 (3.4 ± 0.1)	-1.80*	0.06
ScIce2	1,224	I / E	189	37	0	32 (74.6%)	0.80 ± 0.02	11.2 ± 0.6 (0)	2.34*	0.36
ScVrn1	542	5'UTR/E	198	0	1	2 (97.0%)	0.11 ± 0.03	0.4 ± 0.1 (0)	-0.33	-
Total	9,316			161 (43)	9	339				

[a] E: exon; UTR: untranslated region; I: intron
[b] Failure of amplification in some of the genotypes may be due to the presence of SNPs/Indels in the binding sites of the sequences in some particular genotypes.
[c] Proportion of haplotypes with MAF > 0.05.
[d] Significance levels: *P < 0.05

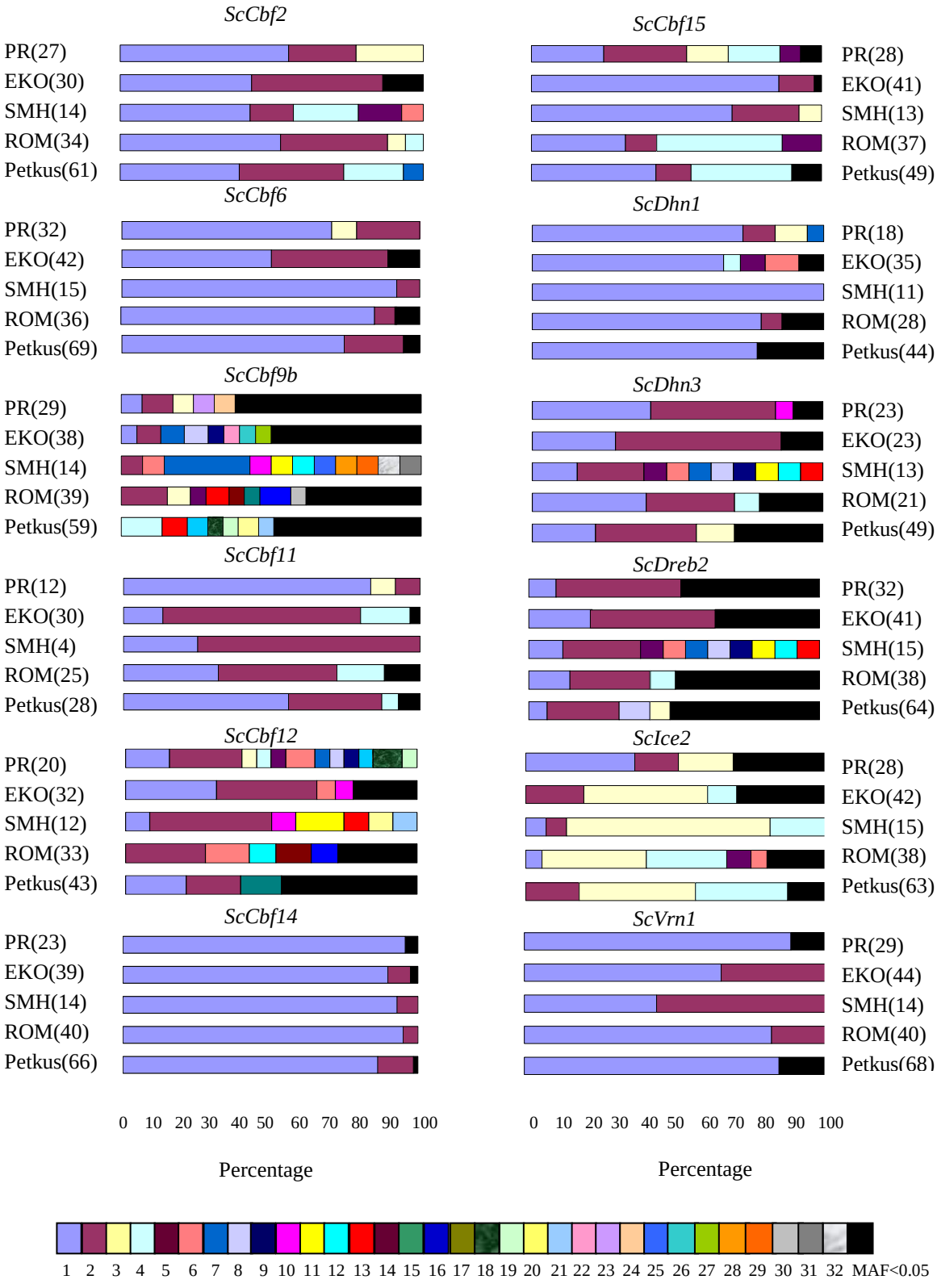


Figure 7 Haplotype frequencies of 12 candidate genes in five rye populations.

The different haplotypes occurring within each gene are represented by different coloured bars (see legend). Haplotypes occurring at a frequency < 0.05 are pooled and shown as black bars. The number of investigated lines in each population is shown in brackets.

3.2.3 Genetic variation within and between populations

PCoA of candidate gene haplotypes revealed large genetic variation within each population and no clustering according to population membership (Figure 8a). The first and second principal co-ordinates explained 10.3% and 9.7% of the total genetic variation, respectively. PCoA of the 37 genome-wide SSRs similarly identified most genetic variation as residing within populations (Figure 8b). However, on the genome-wide level it could differentiate the Petkus population from all Eastern European populations, and the PR population from the other three Eastern European ones. The first and second principal co-ordinates explained 7.3% and 4.1% of the total genetic variation, respectively. AMOVA based on 37 SSR markers and haplotypes of 12 candidate genes both revealed low variation (13.3% and 20.9% of total variation) between populations, but high variation (86.7% and 79.1% of total variation) within populations (Table 9).

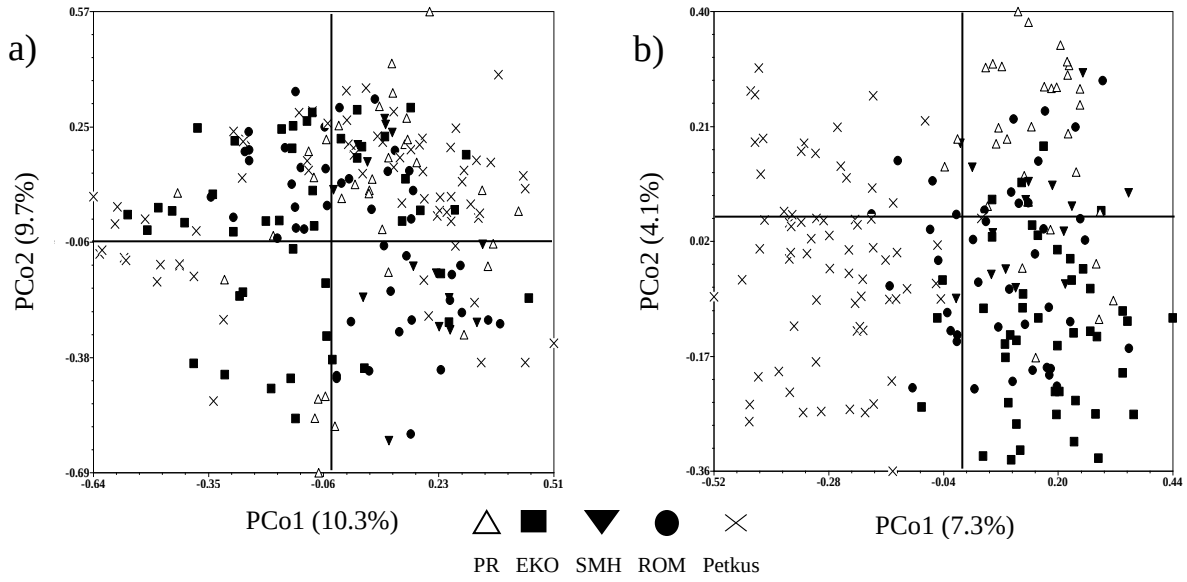


Figure 8 Principal co-ordinate analysis of 201 rye genotypes from five populations based on candidate gene haplotypes (a) and 37 genome-wide SSR markers (b).

PCo1 and PCo2 are the first and second principal co-ordinates and percentages indicate percent variation explained. The names of populations are indicated.

Table 9 Analysis of molecular variance (AMOVA) based on 37 genome-wide SSR markers and haplotypes of twelve candidate genes

Source	Df	Sum of squares	Variance components	Percentage of total genetic variance
SSR markers				
Among populations	4	235,198	1.30***	13.3
Among individuals within population	196	1,801,269	8.50***	86.7
Total	200	2,036,468	9.80	100.0
Candidate genes				
Among populations	4	21255,557	0.80***	20.9
Among individuals within population	196	109367,571	3.01***	79.1
Total	200	130623,128	3.81	100.0

***Indicates significance ($P < 0.001$), obtained from 15,000 permutations

3.3 Population structure and kinship

To avoid false positive results in association analysis, population structure and kinship of 201 genotypes were investigated using 37 genome-wide distributed SSR markers. The STRUCTURE analysis at *k*=2 showed that genotypes from the Petkus were separated from the other populations (Figure 9). At *k*=3 a third cluster (in blue) emerged only in the Eastern European populations comprising 91% of the EKO, 33% of the SMH and 46% of the ROM populations when setting the membership coefficient larger than 0.5. At *k*=4 there are two clear clusters: the cluster in yellow mainly consisting of the PR (Belarus), the cluster in red consisting of the Petkus population (Germany). The remaining genotypes (Poland) have a higher level of admixture. From *k*=5 to 10 most of the new clusters emerged only within the Eastern European populations. Based on the STRUCTURE analysis testing from *k*=2 to 10, the most probable number of populations was *k*=3 where the Ln Pr (X/*k*) (natural log probability of the allele frequency for *k* given clusters) reached a plateau starting from *k*=3. This is an indication of the "true" number of subpopulations. Phylogenetic tree analysis was generally in agreement with the STRUCTURE analysis that the Petkus population is different from the Eastern European populations. Few exceptions were two genotypes from ROM and one genotype from EKO that cluster together with the Petkus population (Figure 9). Additionally, the Petkus population was separated into two groups which could not be seen using the STRUCTURE analysis. Phylogenetic tree analysis did not separate the Eastern European populations clearly; however, one can still see a weak separate clustering of EKO and PR. Both analyses come to a conclusion that the genetic composition of the Petkus population is different from the four Eastern European populations.

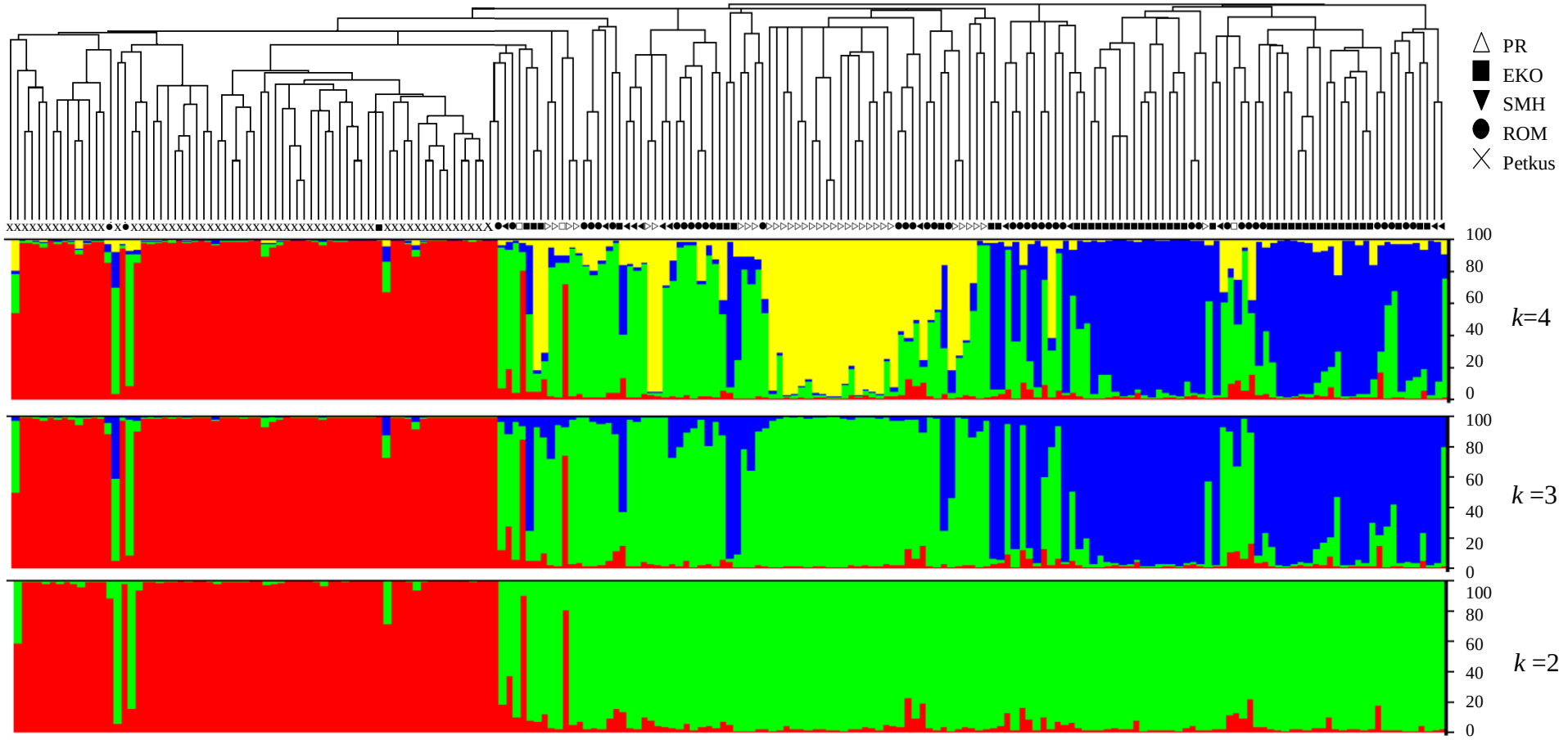


Figure 9 Phylogenetic tree (above) and population structure diagrams (below) based on 37 genome-wide SSR markers.

Each genotype is represented by a tree branch and a vertical bar in the phylogenetic tree and the population structure diagram, respectively. Results of population structure are shown when the numbers of estimated clusters equals *k*=2 to 4. The vertical bar is partitioned into *k* coloured segments that represent the genotype's estimated membership fractions. The legend indicates the symbol and corresponding name of the each population.

The relatedness among the 201 genotypes estimated from the allele-similarity matrix ranged from 0.11 to 1.00 with a mean of 0.37 (Figure 10). Compared to the Eastern European populations, genotypes from Petkus showed a higher relatedness among each other with a mean of 0.53 probably due to more intensive intercross and backcross selections.

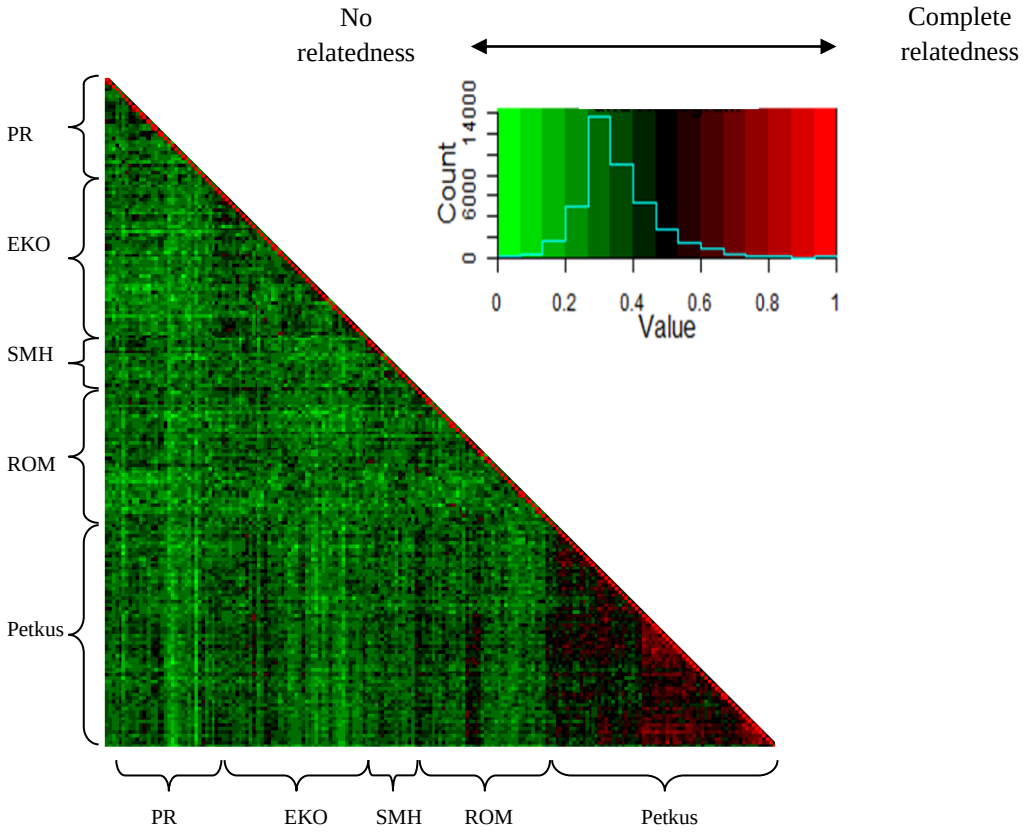


Figure 10 Heatmap of relatedness among the 201 rye genotypes.

Genotypes were sorted according to populations along the x-axis and y-axis. Each grid represents the pair-wise relatedness between genotypes. The colour legend for relatedness with its density distribution is given on the right hand side. Counts are the numbers of pair-wise relatedness between genotypes.

3.4 Decay of linkage disequilibrium

LD decay can vary considerably in different genomic regions due to different recombination rates and selection pressures at different regions of the genome. The mean r^2 for pairs of SNPs within candidate genes ranged from 0.06 to 0.92 (Table 8). The LD patterns varied from gene to gene (Figure 11). Several strong LD blocks were observed in different genes. In *ScCbf11*, two strong LD blocks were observed, one in the interval from SNP1 to SNP12 spanning 99 bp (mean r^2 within LD block = 0.93), and one from SNP17 to SNP27, spanning 243 bp (mean r^2 within LD block = 0.98). In *ScCbf14*, all five SNPs were found to be in high

LD with mean r^2 values 0.92. Estimation of LD in *ScIce2* was performed based on 37 SNPs (mean r^2 = 0.36), all located in the first intron of the gene. There were three strong LD blocks, from SNP1 to SNP18 (block 1), SNP19 to SNP31 (block 2), and SNP32 to SNP37 (block 3), spanning 458 bp, 187 bp, and 61 bp, with a mean r^2 within LD blocks of 0.85, 0.75, and 0.73, respectively. Interestingly, the mean r^2 between blocks 2 and 3 decreased to 0.35, between blocks 1 and 2 further to 0.10, and between blocks 1 and 3 to 0.13. On the contrary, low LD was observed in *ScCbf2* (mean r^2 = 0.13), *ScDhn3* (mean r^2 = 0.25) and in the coding sequence of *ScCbf9b* (mean r^2 = 0.14). The inter-genic LD among the *ScCbf* genes was very low (mean r^2 = 0.05), and only *ScCbf14* showed a slightly higher inter-genic LD (mean r^2 = 0.15) with *ScCbf9b*.

Extent of LD determines mapping resolution and marker density in association studies. The average extent of significant LD pooling all candidate genes and populations together was approximately 520 bp using r^2 = 0.16 as a critical threshold estimated from a separate analysis of 37 unlinked SSR markers (Figure 12). Since the physical distance of each candidate genes are unknown, pairwise comparisons of polymorphic sites were restricted to within candidate genes resulting in 2,194 pairwise comparisons of which almost one third were significant as determined by Fisher's exact test. The average extent of significant LD in individual populations was much smaller because of more stringent threshold values and ranged from 0 bp in the SMH population to approximately 380 bp in the Petkus populations. Locus-specific LD across populations extended from approximately 80 bp in *ScCbf15* to 800 bp in *ScIce2* (Additional figure1). In *ScCbf11*, *ScCbf14*, and *ScDhn1*, mean r^2 remained larger than 0.16 within the 400 bp amplified region. Genome-wide LD based on SSR markers was low with a mean r^2 = 0.01 (Additional table 2). The r^2 was very low and only one pair of markers (RMS1138 versus SCM276) on chromosome 2R exceeded the value of r^2=0.1. On chromosomes 1R and 2R 41.7% and 53.3% of marker pairs showed intra-chromosomal LD at the level of $r^2 > 0.01$ which is a substantially higher proportion than that on the other chromosomes (average = 8.2%).

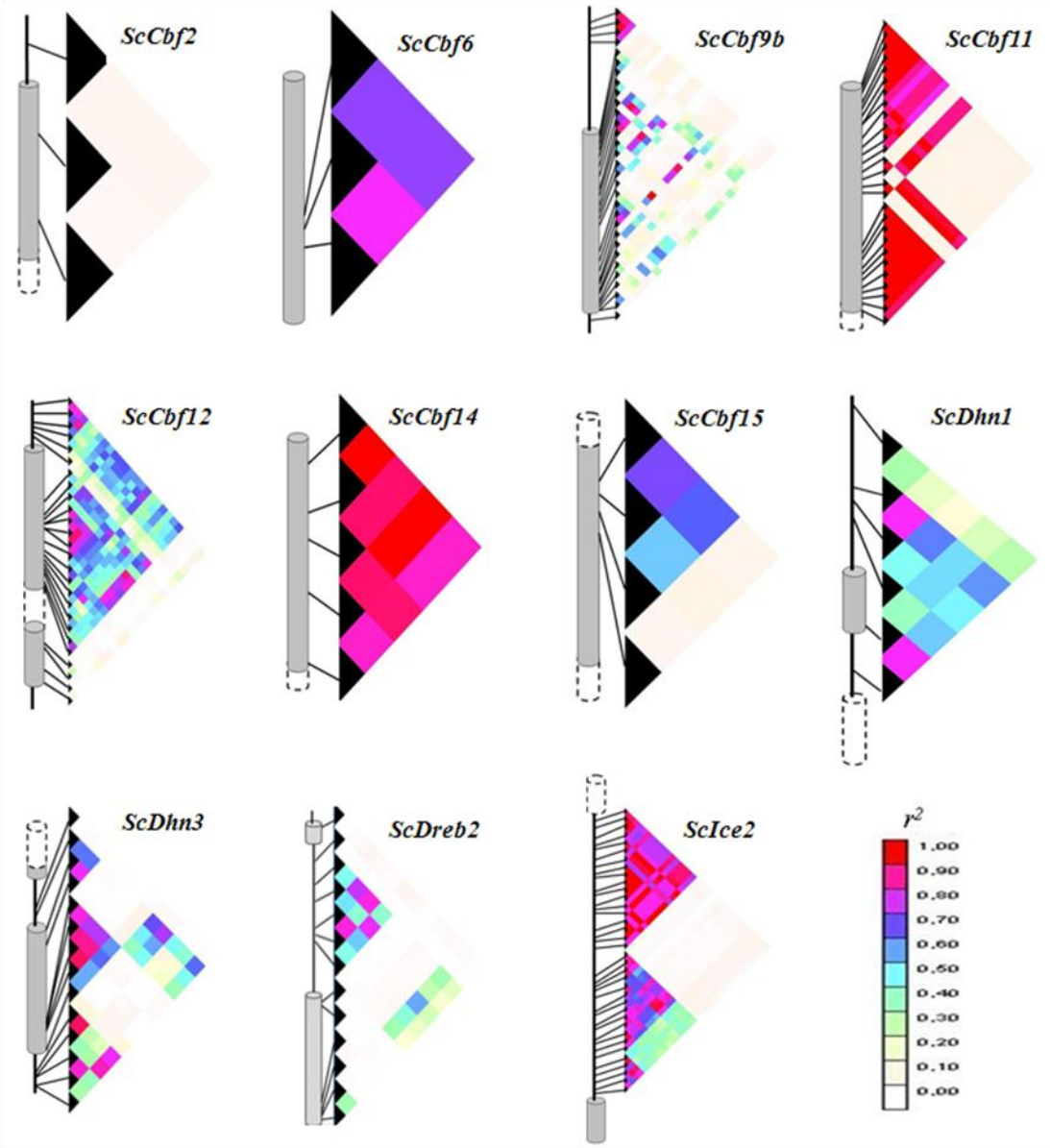


Figure 11 LD heatmaps of twelve candidate genes illustrating pair-wise measurements of LD (r^2) between SNPs.

On the left side of each LD heatmap the gene structure is given whereby exons, and 5‘- or 3‘-flanking regions are represented by grey cylinders and black lines, respectively. White cylinders with dashed lines indicate non-amplified exons. Black triangles represent polymorphic sites starting from “SNP1” on the top of each graph. Each grid represents the strength of LD estimated by r^2 for each pairwise comparison between polymorphic sites with a minor allele frequency (MAF) > 0.05. The colour legend for r^2values is given on the right side. *ScVrn1* was not included due to a lack of pairwise comparisons, since only one Indel was observed.

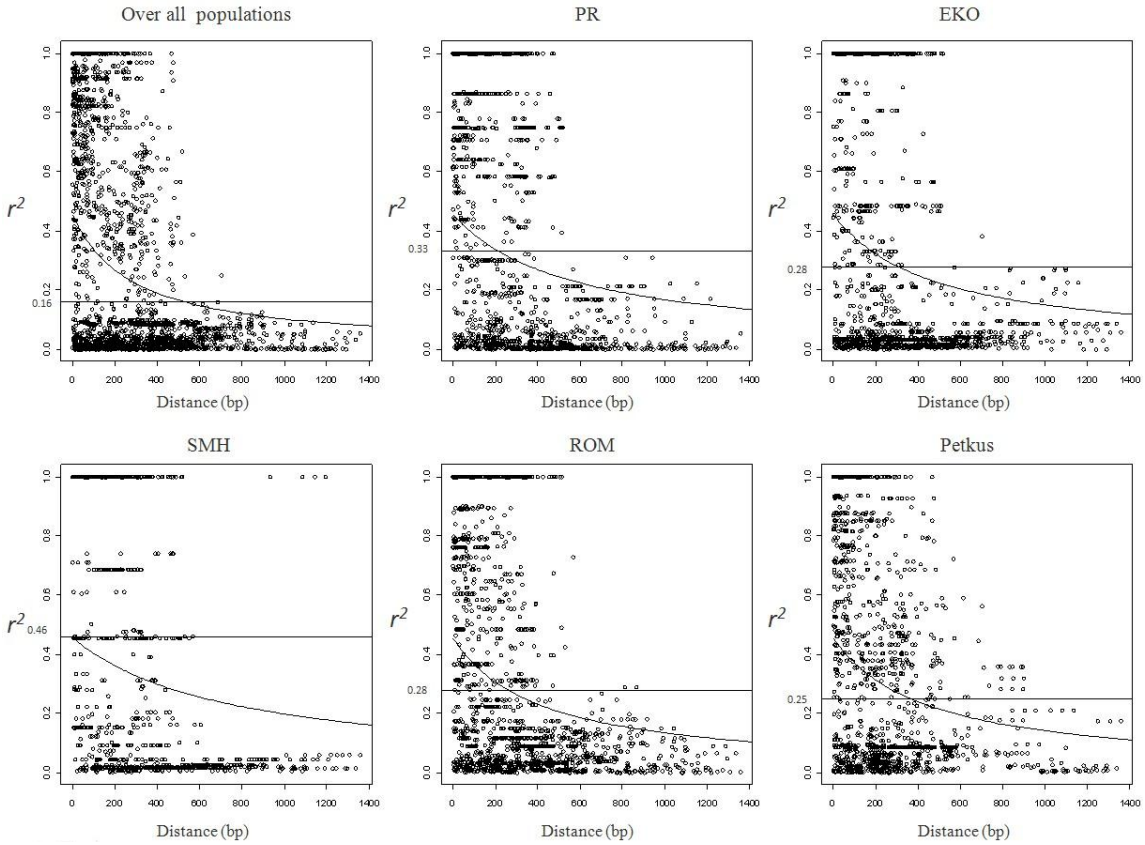


Figure 12 LD decay plots of twelve candidate genes over all populations and in five rye populations.

Non-linear fitting curve from the mutation-recombination-drift model are shown. Thresholds for LD are indicated by a horizontal solid line which varied from 0.16 over all populations to 0.46 in the SMH population (see methods in chapter 2.9).

3.5 Association analyses

3.5.1 SNP-FT associations across three platforms

SNP-FT associations were tested using 170 polymorphisms from twelve candidate genes (161 SNPs and 9 Indels, for convenience further on summarized as "SNPs"). In the controlled platform, 69 statistically significant SNPs were identified among nine genes: *ScCbf2*, *ScCbf9b*, *ScCbf11*, *ScCbf12*, *ScCbf15*, *ScDhn1*, *ScDhn3*, *ScDreb2*, *ScIce2* (all $P < 0.05$; Figure 13). In the semi-controlled platform, 22 statistically significant ($P < 0.05$) SNPs were identified among five genes: *ScCbf2*, *ScCbf11*, *ScCbf12*, *ScCbf15*, and *ScIce2*. In the field platform, 29 statistically significant ($P < 0.05$) SNPs were identified among six genes: *ScCbf9b*, *ScCbf12*, *ScCbf15*, *ScDhn1*, *ScDreb2*, and *ScIce2*. Eighty-four SNPs from nine genes were significantly associated with FT in at least one of the three platforms, and 33

SNPs from six genes were significantly associated with FT in at least two of the three platforms. Across all three phenotyping platforms, two SNPs in *ScCbf15* and one SNP in *ScCbf12* were significantly associated with FT; all of these three SNPs are non-synonymous, causing amino acid replacements. No SNP-FT associations were found for SNPs in *ScCbf6*, *ScCbf14*, and *ScVrn1*. Full information on SNP-FT associations for all platforms can be found in Additional table 3.

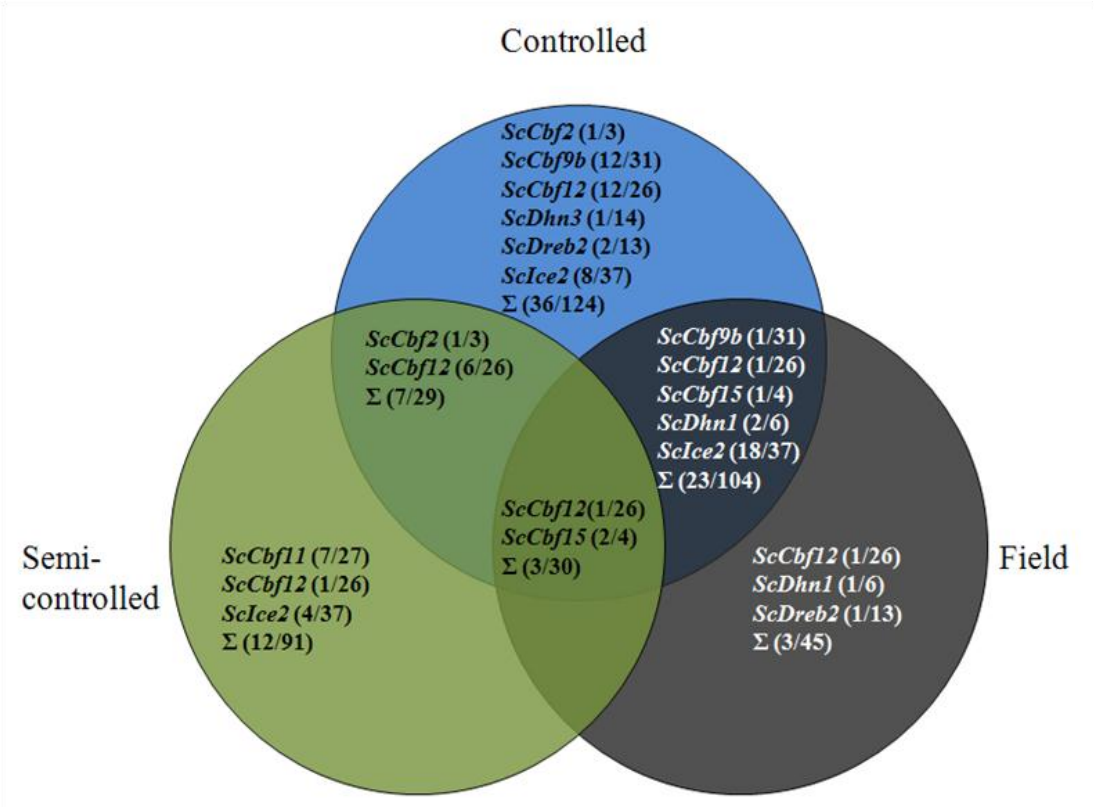


Figure 13 Venn diagram showing SNPs from candidate genes significantly ($P < 0.05$) associated with frost tolerance in one, two, and three phenotyping platforms.

The first and second numbers in brackets are the number of significant SNPs and total number of SNPs in each candidate gene, respectively.

Over all platforms the average percentage of significant SNP-FT associations was higher in the promoter and intron regions than in exons and 3'UTR (Figure 14). On average, 33.3% of significant SNP-FT associations were observed in promoters, 35.5% in introns, 17.2% in exons, 2.9% in 5'UTRs and 11.1% in 3'UTRs, suggesting that the regulatory elements might play an important role in the response to frost stress. It is worth pointing out that the vast majority (96%) of the intronic SNPs significantly associated with FT were located in the first

intron of *ScIce2* with high LD and consequently lead to high percentages of significant SNP-FT associations in this region. Comparing the three platforms, the controlled platform had the highest percentages of significant SNP-FT associations except for the 3'UTR. Surprisingly, the percentages of significant SNP-FT associations differed only slightly between synonymous and non-synonymous SNPs.

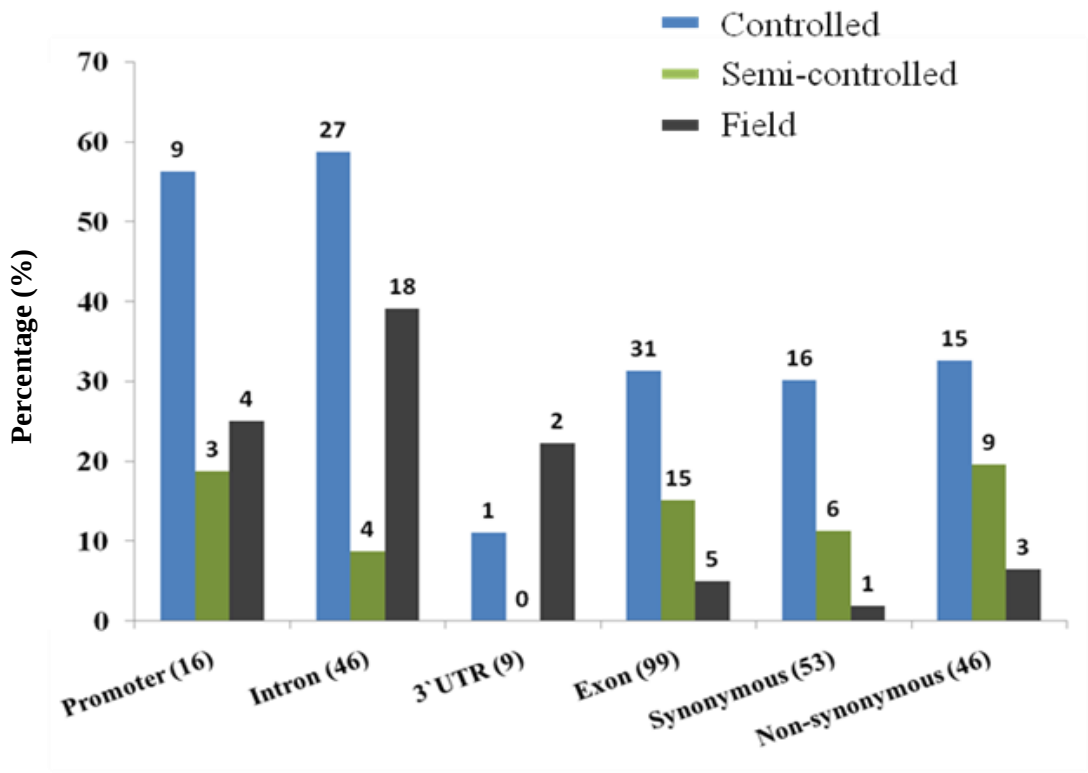


Figure 14 The percentages of significant ($P < 0.05$) SNP-FT associations sorted according to gene regions and types of SNPs in the exons

Histograms show the percentages of significant SNP-FT associations in all the SNPs belonging to the specific gene regions or types of SNPs in the exons. Numbers in brackets on the x-axis are the total number of SNPs in this gene region. The numbers on top of each bar are the number of SNPs significantly associated with FT.

Allelic effects (β_{SNP}) of the 170 SNPs studied were relatively low, ranging from -0.43 to 0.32 for recovery scores in the controlled platform, -2.17% to 2.44% for % plants with undamaged leaves in the semi-controlled platform, and -3.66% to 4.30% for % survival in the field platforms (Figure 15). Among all significant SNPs, the smallest allelic effects (β_{SNP}) in the controlled, semi-controlled and field platform were 0.16%, 1.96%, and 2.01% respectively. 45.5% of all significant SNPs found in at least one platform had positive allelic effects, indicating the non-reference allele conveyed superior FT to the reference allele. The largest positive β_{SNP} among the 170 SNPs in the field platform was observed for SNP7 in *ScIce2*

(β_{SNP} = 4.30). This favorable allele was present predominantly in the PR population (55.2%), and occurred at much lower frequency in the other four populations (EKO: 4.7%, SMH: 6.7%, ROM: 7.1%, and Petkus: 0%). The proportion of genetic variation explained by individual SNPs ranged from 0% to 27.9% with a median of 0.4% in the controlled platform, from 0% to 25.6% with a median of 1.2% in the semi-controlled platform, and from 0% to 28.9% with a median of 2.0%in the field platform (Figure 16). These distributions were highly concentrated near 0.

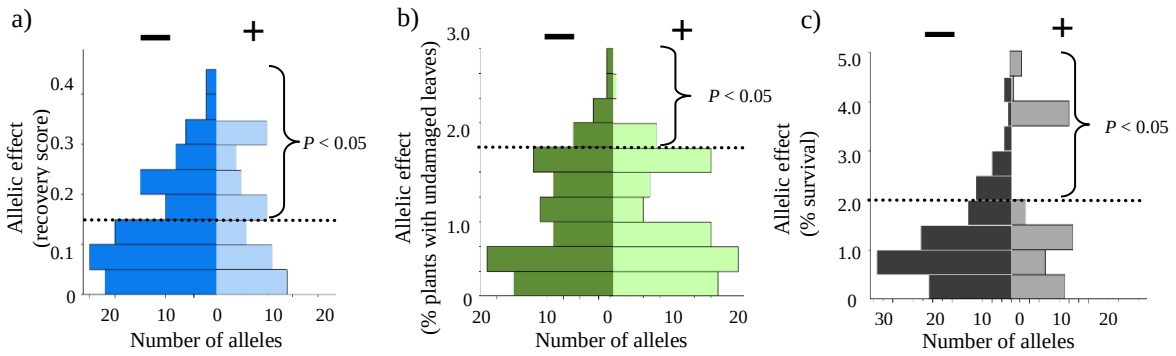


Figure 15 Distribution of allelic effects (β_{SNP}) of SNP-FT associations in a) controlled, b) semi-controlled, and c) field platforms.

The left and right hand side bars in a), b) and c) represent alleles with negative (-) and positive (+) effects relative to the Lo152 reference allele, respectively. The significance threshold ($P < 0.05$) for each platform is indicated by a dashed line.

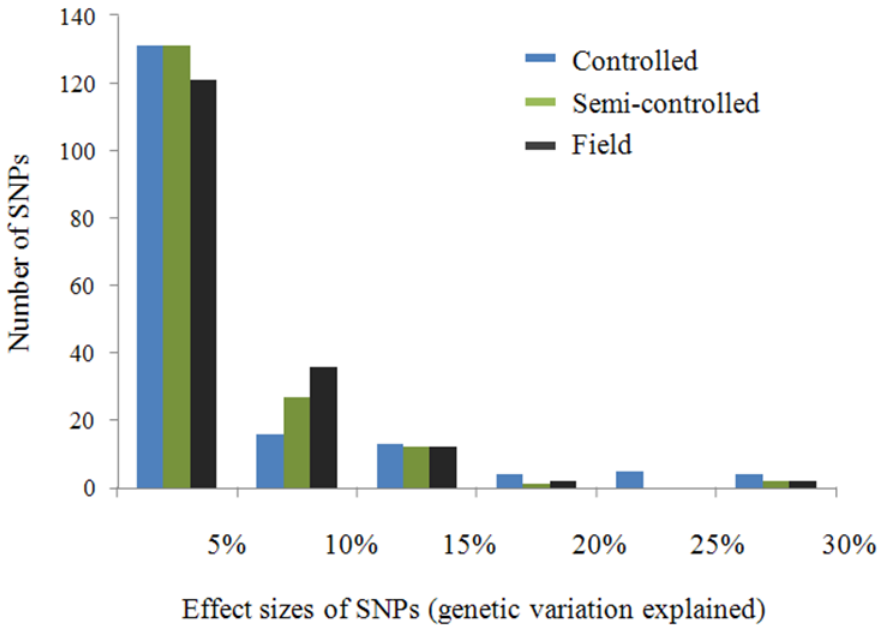


Figure 16 Distributions of effect sizes of SNPs (genetic variation explained by individual SNPs) in three phenotyping platforms.

Empirical correlations of the SNP-FT association results, in terms of *t* values, between the three phenotyping platforms were moderate to low (Figure 17). The highest correlation coefficient was observed between the controlled and semi-controlled platforms with $r = 0.59$, followed by correlations between the controlled and field platforms with $r = 0.54$, and the semi-controlled and field platforms with $r = 0.24$. When correlations were restricted to the significant SNPs, slightly higher correlation coefficients were observed with $r = 0.64$ between the controlled and semi-controlled platforms, $r = 0.66$ between the controlled and field platforms, and $r = 0.34$ between the semi-controlled and field platforms.

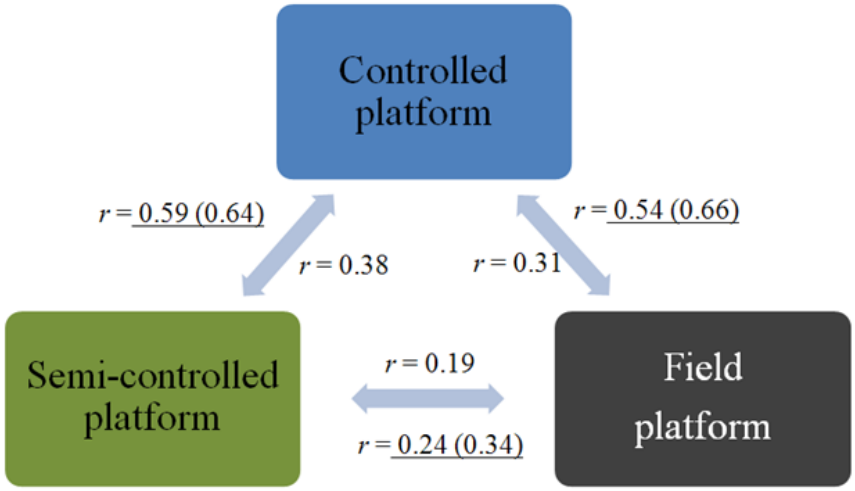


Figure 17 Correlation coefficients of the three different phenotyping platforms.

The underlined numbers were based on empirical correlations of the *t*-values from all 170 SNPs and 84 (in brackets) SNPs significantly associated with FT in at least one of the three platforms, respectively. The inner numbers are based on the phenotypic adjusted entry means.

3.5.2 Haplotype-FT associations

Haplotype-FT associations were performed using 30 haplotypes (MAF > 5%) in eleven candidate genes. Because only one haplotype in *ScDhn1* had a MAF > 5%, *ScDhn1* was excluded from further analysis. Large numbers of rare haplotypes (MAF < 5%) were found in *ScCbf9b* (N=62) and *ScCbf12* (N=22) resulting in large numbers of missing genotypes (87.9% and 61.3%) for the association analysis. Haplotypes 2, 3, and 4 in *ScCbf2* were significantly ($P < 0.05$) associated with FT in the controlled platform. Haplotype 3 in *ScCbf2* which had a positive effect on FT, was present mainly in the PR population (20.7%), whereas occurred in much lower frequencies in the other four populations (0.0% in EKO, 0.0% in SMH, 5.1% in ROM, and 0.0% in Petkus, Figure 7). For haplotypes 1 and 2 in *ScCbf15* and haplotype 1 in

ScIce2, significant associations ($P < 0.05$) were found across two and three platforms, respectively (Table 10). Haplotype effects (β_{Hap}) were relatively low and comparable to the allelic effects (β_{SNP}) ranging from -0.31 to 0.49 (recovery score), -1.71% to 2.74% (% plants with undamaged leaves), and -3.32% to 3.80% (% survival) in the controlled, semi-controlled and field platforms, respectively. The highest positive effect on survival rate was observed for haplotype 1 of *ScIce2* in the field platform, implicating this haplotype as the best candidate with superior FT. This favorable haplotype was present mainly in the PR population (35.7%), occurring in much lower frequencies in the other four populations (0.0% in EKO, 6.7% in SMH, 5.3% in ROM, and 0.0% in Petkus (Figure 7). The proportion of genetic variation explained by all haplotypes ranged from 0% to 25.7% with a median of 1.6% in the controlled platform, from 0% to 17.6% with a median of 1.4% in the semi-controlled platform, and from 0% to 9.3% with a median of 4.8% in the field platform.

Table 10 Summary of haplotypes significantly associated with FT in at least one platform, their haplotype effects, and percent genetic variation explained by the haplotypes

Candidate gene	Name of haplotype [a]	Controlled (recovery score 0-5) [b]			Semi-controlled (% plants with undamaged leaves)			Field (% survival)		
		***P*-value [c]**	**β_{Hap}**	**% genetic variation explained**	***P*-value**	**β_{Hap}**	**% genetic variation explained**	***P*-value**	**β_{Hap}**	**% genetic variation explained**
ScCbf2	Overall[d]	**<0.001**	-	25.7	0.21	-	16.3	0.40	-	5.0
	2	**0.04**	-0.11	-	0.51	-0.51	-	0.73	-0.51	-
	3	**<0.001**	0.49	-	0.19	1.36	-	0.12	3.32	-
	4	**<0.001**	-0.31	-	0.12	-1.43	-	0.74	0.57	-
ScCbf15	Overall	**<0.01**	-	0.6	0.09	-	17.6	0.09	-	4.4
	1	**<0.01**	-0.22	-	**0.04**	-1.69	-	0.06	-3.32	-
	2	**<0.01**	-0.21	-	0.13	-0.92	-	**0.04**	-2.59	-
ScIce2	Overall	**0.04**	-	4.8	**0.02**	-	13.3	0.13	-	8.1
	1	**<0.01**	0.29	-	**<0.01**	2.74	-	**0.02**	3.47	-

[a] Haplotypes with minor allele frequency (MAF) > 5%
[b] 0: completely died. 1: little sign of life. 2: intensive damage. 3: moderate damage. 4: small damage. 5: no damage
[c] *P*-values < 0.05 are printed in bold
[d] All haplotypes (MAF > 5%) within a candidate gene

3.5.3 Complex epistasis contributing to frost tolerance

Epistasis, or gene × gene interaction, is an important context-dependent genetic effect. Out of all 55 possible gene × gene interactions (eleven genes with more than two haplotypes with MAF > 5%), 30 were tested since the rest were dominated by a single haplotype combination (> 95%). Among the 30 possible gene × gene interactions, eleven, six, and one were significantly ($P < 0.05$) associated with FT in the controlled, semi-controlled, and field platforms, respectively. *ScCbf15* × *ScCbf6*, *ScCbf15* × *ScVrn1*, *ScDhn3* × *ScDreb2*, and *ScDhn3* × *ScVrn1* were significantly associated with FT across two platforms, none were significantly associated with FT across all three platforms (Figure 18). *ScVrn1*was involved in eight gene × gene interactions which was the largest number despite the fact that it was not significantly associated with FT in any platform in the statistical model without interaction term. Two gene × gene interactions occurred between members of the *Cbf* gene family: *ScCbf6* × *ScCbf15* and *ScCbf11*× *ScCbf14*.

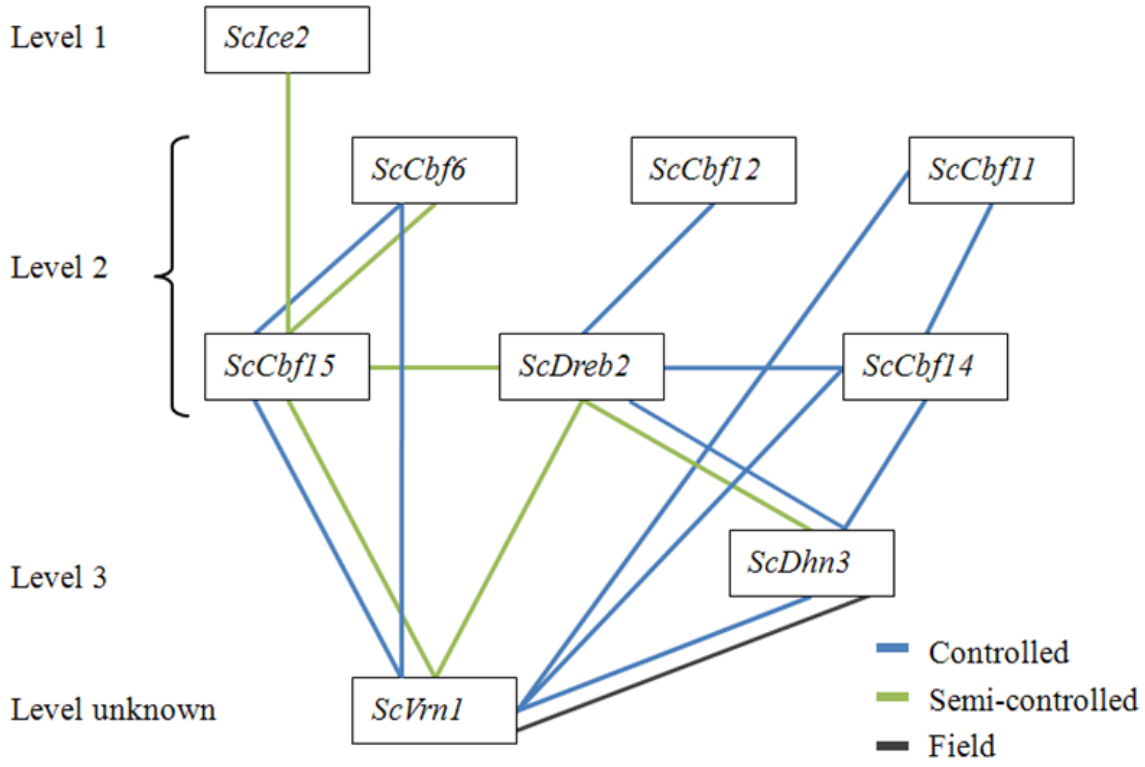


Figure 18 Significant ($P < 0.05$) gene × gene interactions for frost tolerance in three phenotyping platforms.

Candidate genes are sorted in three known and one unknown level according to the frost responsive cascade (Yamaguchi-Shinozaki and Shinozaki 2006).

3.5.4 SNP by environment interactions

SNP by environment interaction is another important context-dependent genetic effect. In the controlled platform, 95 SNPs with significant ($P < 0.05$) SNP by environment (temperature and year) interactions were identified among all candidate genes except for *ScVrn1* (Figure 19). In the semi-controlled platform, 70 SNPs with significant SNP by environment (year) interactions were identified among eight genes: *ScCbf2*, *ScCbf9b*, *ScCbf11*, *ScCbf12*, *ScCbf15*, *ScDhn1*, *ScDhn3* and *ScIce2*. In the field platform, 58 SNPs with significant SNP by environment (field location) interactions were identified among nine genes: *ScCbf6*, *ScCbf9b*, *ScCbf12*, *ScCbf14*, *ScCbf15*, *ScDhn1*, *ScDhn3*, *ScDreb2* and *ScIce2*. There are 147 SNPs from eleven genes (except for *ScVrn1*) with significant SNP by environment interactions in at least one of the three platforms, and 55 SNPs from eleven genes (except for *ScVrn1*) in at least two of the three platforms. Across all three phenotyping platforms, 21 SNPs from *ScCbf15*, *ScDhn1*, *ScDhn3* and *ScIce2* with significant SNP by environment interactions were found. Full information on SNP × environment interactions for all platforms can be found in additional table 3.

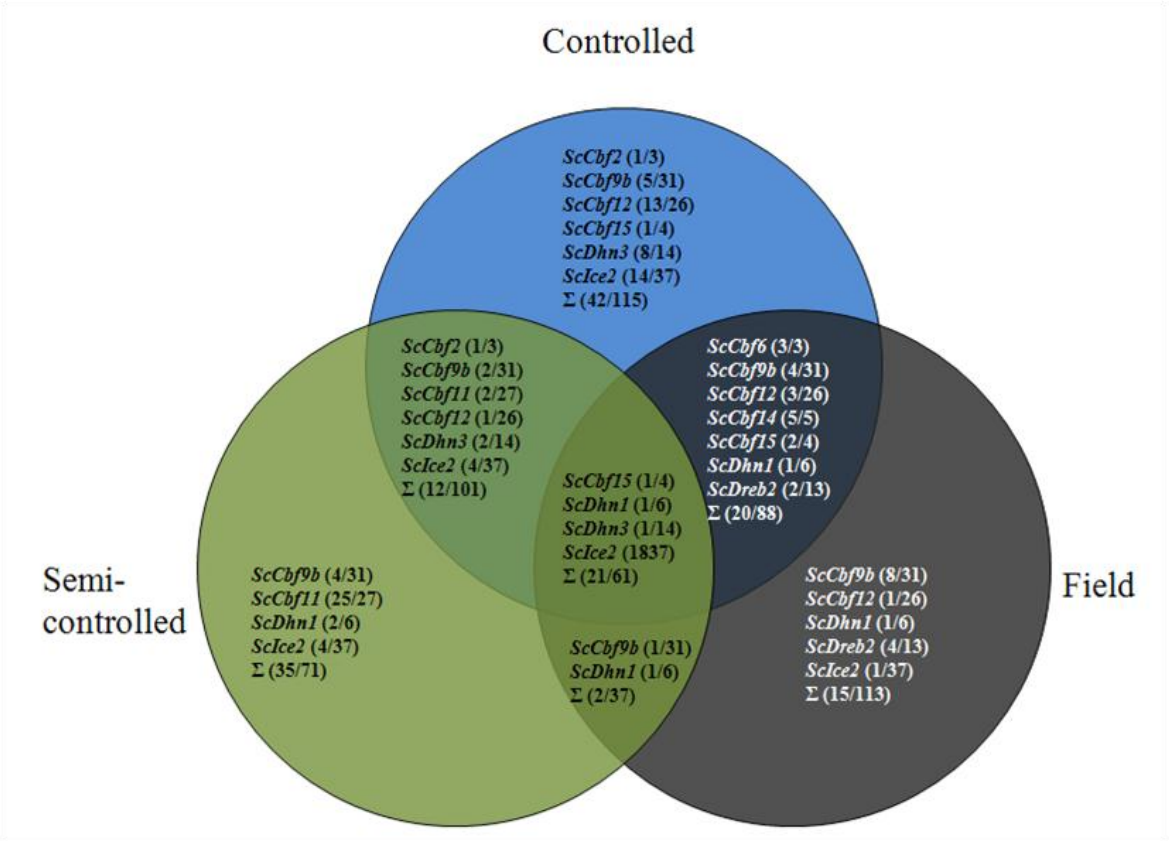


Figure 19 Venn diagram showing SNPs from candidate genes with significant ($P < 0.05$) SNP × environment interaction in one, two, or three phenotyping platforms.

The first and second numbers in brackets are the number of significant SNP × environment interactions and total number of SNPs in each candidate gene.

4. Discussion

4.1 A diverse germplasm collection for winter rye association studies

The choice of germplasm representing phenotypic and genetic diversity of the examined species is important in association studies. Theoretically, diverse germplasm including worldwide cultivars, landraces, and possibly even wild species should be used to fully exploit ancestral recombination and mutation events. In this study germplasm was restricted to Middle and Eastern European rye populations due to the following two reasons. Firstly, the five Middle and Eastern European populations are good representatives of the worldwide germplasm because 81% of rye production in the world in 2009 was located in Middle and Eastern Europe. Secondly, the primary aim of this study was to improve FT of the Middle European Petkus population by identifying favorable alleles from Eastern European germplasm which might have undergone strong selection pressure for FT. Thus, germplasm from countries with mild winters such as Spain, Turkey, and Iran is not expected to possess such favorable alleles due to lack of selection pressure. Phenotypic variation in 12 environments from three phenotyping platforms was high and mainly attributed to genotypes. Genetic diversity was also high based on both sequencing data of candidate genes and SSRs.

4.1.1 Comparing genetic diversity of the five winter rye populations to other species

Assessment of genetic diversity based on genome-wide SSRs and locus-specific candidate genes are complementary investigations, the former providing a global view of the rye genome and the latter focussing on genes involved in the frost tolerance network. Nucleotide and haplotype diversity of twelve candidate genes were investigated in the five winter rye populations from Middle and Eastern Europe. SNP frequency observed in a total of around 10 kb DNA sequence was on average one SNP every 52 bp and the average nucleotide diversity (π) ranged from 0.4×10^{-3} to 14.5×10^{-3} with an average value of $\pi = 5.6\times10^{-3}$. These values are as high as those reported in maize landraces, where one study reported a rate of one SNP per 62 bp, a range of π from 0.1×10^{-3} to 13.3×10^{-3} and an average value of π equal to 4.0×10^{-3} (Yamasaki et al. 2008). Some studies have suggested that comparisons among different species should be restricted to homologous genes (Krutovsky and Neale 2005). Nucleotide diversities of three *Cbf* homologs in 34 *Arabidopsis* ecotypes (*AtCbf1, AtCbf2* and *AtCbf3*) ranged from $\pi = 2.6\times10^{-3}$ to 6.9×10^{-3} (Lin et al. 2008), a smaller range compared to this study ($\pi = 1.5\times10^{-3}$ to 14.5×10^{-3}), which is likely due to the different mating system (Nei 1987). Observed haplotype diversities of *HvCbf9b* in *Hordeum spontaneum*, old cultivars,

and modern cultivars of *H. vulgare* were 0.48, 0.18, and 0.06, respectively, which is much lower than that of *ScCbf9b* (*Hd* = 0.98 ± 0.03) in this study (Fricano et al. 2009). Another example also showed higher sequence diversity of rye compared to barley using 14 amplicons of rye derived from barley expressed sequence tags (ESTs) putatively involved in biotic and abiotic stress tolerance (Varshney et al. 2007). This again indicates that the mating system plays an important role on genetic diversity between different species. To summarize, genetic diversity in this rye germplasm is high and therefore suitable for association studies.

Natural selection leaves its traces in the pattern of nucleotide polymorphism in the genes. Strong directional selection on the loci responsible for some kind of abiotic stress tolerance should reduce diversity within locally adapted populations due to increases in frequency of alleles conferring adaptation. However, no reduction of genetic diversity was observed in the Eastern European populations compared to the Petkus population based on FT candidate genes. One possible explanation is that at the time when selection of FT took place, winters in Germany, the provenance of the Petkus population, were harsh enough to form a similar selection pressure on the Petkus population compared to Eastern European populations under Eastern European winters. It must be stated however, that Petkus is the only representative for the Middle European rye populations in this study and thus conclusions on population differences must be limited to the Petkus population. Another reason could be that FT is a complex quantitative trait involving large gene networks comprising individual genes contributing only small effects, thereby making it difficult to detect selection signatures, such as reduction of genetic diversity in candidate genes.

On the contrary, genome-wide assessment of diversity using SSR markers revealed a higher genetic diversity for the Eastern European populations PR, EKO, SMH, and ROM compared to the Middle European Petkus population. One reason for this finding might be a bottleneck effect due to a higher selection pressure in the Petkus population, whereby it could be assumed that many “unfavourable” minor alleles were eliminated to pave the way for plants with desirable traits. The Petkus population, one of the two heterotic groups in rye, has systematically been improved by more than five cycles of full sib recurrent selection for yield, thousand-grain weight, resistance to lodging, self fertility, and disease resistance. A reduction in allele diversity of SSR markers due to hitchhiking with linked loci which were targets of selection is therefore probable. The reduction of genetic diversity due to human-induced selection has been well documented in barley and maize (Fricano et al. 2009; Kilian et al. 2006; Tian et al. 2009). By contrast, the Eastern European populations experienced a lower

selection pressure by mass or half sib selection in the breeding programs where introgression of foreign material was common in order to keep genetic variability on a high level. Therefore, the Eastern European populations can be used for broadening the genetic diversity of the elite breeding materials such as the Petkus population.

4.1.2 Genetic differentiation between the Petkus and Eastern European populations

Based on the STRUCTURE analysis testing from *k*=2 to *k*=10, the most probable number of populations was *k*=3 where the Middle European Petkus population was clearly separated from the Eastern European populations. This conclusion is based on three reasons: Firstly, the Ln Pr (X/K), which is the natural log probability of the data (allele frequency) for *k* given clusters that are present in the data, reached a plateau starting from *k*=3. This is an indication of the "true" number of subpopulations by Pritchard et al. (2000b). Secondly, *k* =3 can be explained by the cultivation history of rye. According to the most recent archaeological evidences, rye was first cultivated at Abu Hureyra, in modern Syria (Hillman et al. 2001; Murphy 2007) and later spread to Russia and Belarus via Caucasia and then into Poland and Germany (Bushuk 2001; Salamini et al. 2002). This is consistent with the observation of this study that EKO, SMH, and ROM (from Poland) have membership fractions from PR (from Belarus). The Petkus population, which is elite material from KWS-LOCHOW GMBH, has gone through intensive and effective selections for e.g. yield, resistance to lodging, and disease resistance. Therefore, the Middle European Petkus population is clearly separated from the Eastern European populations. This was further confirmed by phylogenetic tree analysis and PCoA based on SSR markers.

4.1.3 Selection pattern and decline of LD

LD results from the interplay of many factors. Selection, which causes locus-specific bottlenecks, is one of the factors that increase LD within and between alleles of selected loci. In the present germplasm, *ScCbf15* and *ScIce2* showed significant positive values of Tajima's *D* (2.14 and 2.34, respectively; $P < 0.05$) over all populations, indicating balancing selection, whereby genotypes carrying alleles with intermediate frequency are favored. Positive Tajima's *D* values can also be observed if a population was formed from a recent admixture of two different populations, which cannot be excluded in this study. *ScDhn1* and *ScDreb2* showed a significantly negative value of Tajima's *D* ($P < 0.05$), indicating purifying selection, whereby an excess of polymorphisms with low frequencies was observed. Interestingly, *Dhn1* in Scots pine has also been described as subject to positive selection

(Wachowiak et al. 2009), implying that *Dhn1* is possibly a target of selection in different species. In this study, a large variation of mean r^2 in seven *Cbf* genes (0.13 to 0.92) was observed, indicating that the *Cbf* gene family has probably undergone diverse selection history. The extent of LD across all twelve candidate genes and over all rye populations was approximately 520 bp. This rapid decay of LD could be expected, because compared to self-pollinated species, cross-pollinated rye has a higher effective recombination rate (Flint-Garcia et al. 2003). In contrast to self-pollinated species, where LD extends up to 212 kb in cultivated barley(Caldwell et al. 2006b), a comparable fast LD decay has been reported in other cross-pollinated species, including douglas fir, maize, and ryegrass (Krutovsky and Neale 2005; Tenaillon et al. 2001; Xing et al. 2007). Pairwise LD measured by r^2 based on SSRs was very low (mean $r^2 = 0.01$), which was expected since the 37 SSRs have an average marker interval of 21 cM according to the integrated consensus map of Gustafson et al. (2009). Knowledge about LD can give hints for the marker density that is required for GWAS. The rapid decay of LD in rye promises a high resolution mapping in GWAS. A challenge, however, is that a huge number of markers is required for covering the whole genome encompassing more than 8,000 Mb. The exact number required for GWAS in rye is difficult to estimate with this study since LD decay could vary considerably from locus to locus due to different recombination rates and selection pressures at different regions of the genome. A more complete picture of LD extent in rye might be obtained using genome-wide distributed SNP markers (Haseneyer et al. 2011).

4.2 Association analyses

4.2.1 Biological implications of candidate genes with significant associations

Statistically significant SNP-FT associations were identified in nine out of twelve candidate genes hypothesized to be involved in the frost responsive network among which the transcription factor *Ice2* is one of the key factors. The function of *Ice2* was characterized both in wheat and *Arabidopsis* (Badawi et al. 2008; Fursova et al. 2009). Over-expression of *TaIce2* and *AtIce2* in transgenic *Arabidopsis* plants resulted in increased FT of transgenic plants and was associated with higher expression levels of the *Cbf* gene family. Using electrophoresis mobility shift assays, Badawi et al. (2008) further showed that *TaIce2* binds to the promoter region of *TaCbf9*. Unfortunately, no interaction between *ScIce2* and *ScCbf9b* could be observed in this study probably due to the large number of rare haplotypes (MAF < 5%) in *ScCbf9b* resulting in many missing genotypes (87.9%) and thus in insufficient

statistical power to identify gene × gene interaction. However, in the rye homolog *ScIce2* we detected 30 out of 37 SNPs in high LD (average r^2 = 0.85) which were significantly associated with FT. These results support the findings of expression studies that *Ice2* is one of the key elements in the frost responsive network. Given that these 30 SNPs are all located in the first intron of the gene, they are unlikely to be functional. However, it is possible that they are in LD with functional polymorphisms located in the coding sequence (CDS) of the gene which we have not investigated due to a lack of rye sequences in GenBank for primer design. The favorable allele of SNP7 in *ScIce2* had a relatively large allelic effect on FT in the controlled and field platforms when compared to other SNPs in this study. This allele was present predominantly in the PR population while entirely absent in the Petkus population. Thus, this SNP might facilitate marker-assisted backcrossing to introgress favorable genomic regions into the Petkus population, thereby improving FT of current breeding materials.

The *Cbf* gene family, regulated by *Ice2*, belongs to the family of *APETALA2* (*AP2*) transcription factors, some of which (except *Cbf11*) are closely linked in cereals and map to the FT locus *Fr2* on homoeologous group 5 of barley and wheat spanning approximately 0.8 cM in the genetic maps (Baga et al. 2007; Francia et al. 2007; Knox et al. 2008). The order of *Cbf* genes in the genetic map is consistent in both species and they share high sequence similarity (Galiba et al. 2009). Expression studies have revealed that the *Cbf* gene family is involved in the frost responsive network in diverse species (Campoli et al. 2009; Hannah et al. 2005; Stockinger et al. 2007). In this study, seven *Cbf* genes were investigated and statistically significant associations were found in at least one platform for *ScCbf2*, *ScCbf9b*, *ScCbf11*, *ScCbf12*, and *ScCbf15* but not for *ScCbf6* and *ScCbf14*. This confirms previous studies that not all members of the *Cbf* gene family are involved in the frost responsive network (Campoli et al. 2009; Stockinger et al. 2007). In *ScCbf2*, a 200 bp Indel was highly associated ($P = 6.27e^{-5}$) with FT in the controlled platform and explained a high proportion of the genetic variation in the controlled (25.7%) and semi-controlled (16.3%) platforms. It is noteworthy that this 200 bp Indel in the promoter of *ScCbf2* contained two MYB and one MYC *cis*-elements. In wheat the presence of MYB and MYC elements has been shown to affect the binding specificity of TaIce41 (wheat homolog of ScIce2) and consequently the expression level of the *TaCbf* gene family (Badawi et al. 2008). Expression studies are needed to investigate the effect of multiple binding sites for ScIce2 in *Cbf* gene promoters on the expression level of *Cbf* genes. A study in *Triticum monococcum* suggested that polymorphisms in *TmCbf12*, *TmCbf14*, and *TmCbf15* are the most likely explanation for observed differences in FT (Knox et al. 2008). Among the five significantly associated *Cbf*

genes in our study, SNP17 in *ScCbf12*, as well as SNP1 and SNP2 in *ScCbf15* were significantly associated with FT across all three platforms. Given that these three SNPs are all non-synonymous, leading to amino acid exchanges in the CDS of their respective genes, they are good candidates for functional genetic studies. In a recent candidate gene-based association study, Fricano et al. (2009) found two SNPs located in the 3'-untranslated region of *HvCbf14* significantly associated with FT in barley. The 3'-untranslated region of *ScCbf14* was not sequenced in this study; it would be interesting to sequence this region to investigate whether it also contains SNPs significantly associated with FT in rye as well. However, members of the *Cbf* gene family are not the only key factors in the frost responsive network (Fowler and Thomashow 2002; McKhann et al. 2008). Hannah et al. (2005) reported that 45% of the *Arabidopsis* transcriptome was cold responsive, but only 33% of the cold responsive transcriptome belonged to the *Cbf* regulon. In a study of wheat, Monroy et al. (2007) reported that at least one-third of the genes induced by cold did not belong to the *Cbf* regulon. The transcription factor *AtHOS9*, which encodes a putative homeobox protein, has been shown to contribute to the regulation of FT in *Arabidopsis* independently of the *Cbf* regulon (Zhu et al. 2004). Thus, extending research to the analysis of more candidate genes of the frost responsive network in this dataset would certainly be worthwhile.

The dehydration-responsive element binding gene, *Dreb2*, another member of the *AP2* transcription factor family, has been isolated and characterized in several crop species such as wheat, barley, maize, and rice (Dubouzet et al. 2003; Egawa et al. 2006; Qin et al. 2007; Xue and Loveridge 2004). Similar to *Cbf* genes, *Dreb2* can specifically bind to DRE/CRT *cis*-elements of the stress-inducible target genes, albeit primarily under drought rather than cold/frost stress (Liu et al. 1998). However, it is not surprising that *Dreb2* can also be induced by cold/frost as shown by recent studies in wheat and maize since both drought and cold/frost stresses lead to dehydration of cells (Egawa et al. 2006; Qin et al. 2007). In this study, three SNPs in *ScDreb2* were significantly associated with FT supporting the hypothesis that *Dreb2* in rye is not only involved in drought response but also in frost response.

The dehydrin genes, part of the *COR* gene family, are regulated by the *Cbf* gene family and the *Dreb2* gene via the *cis*-element DRE/CRT present in the promoter region of *COR* genes (Yamaguchi-Shinozaki and Shinozaki 2006). Transcripts of *HvDhn1*, *HvDhn3* and other *HvDhn* genes were detected under frost stress in barley (Zhu et al. 2000). We detected SNP2 and SNP3 in high LD ($r^2 = 0.93$) in the promoter region of *ScDhn1* and SNP3 of *ScDhn3* with significant associations with FT in the controlled platform. These SNPs might serve as

variants which affect the binding specificity of the *Cbf* gene family. This study confirmed results from expression studies in barley, wheat and *Arabidopsis* and suggests that *Ice2*, the *Cbf* family, *Dreb2*, and the *Dhn* gene family are involved in the frost responsive network in rye as well.

4.2.2 Effect sizes of SNP-FT associations

Effect sizes of markers, commonly expressed as percentage of the genetic variance explained by markers, are of primary interest in association studies since they are the main factors that determine the effectiveness of subsequent marker assisted-selection processes. Studies on the genetic architecture of quantitative traits have attracted a lot of interest and became a hot topic in recent years (Buckler et al. 2009; Flint and Mackay 2009; Ingvarsson and Street 2010). Small allelic effects were found in maize flowering time where the largest effect of a QTL allele was only 0.4 days as measured by anthesis-silking interval (Buckler et al. 2009). A recent review summarizing association studies in 15 different plant species implicated that the effect size of QTL depends largely on the phenotypic traits, species, and types of variants (Ingvarsson and Street 2010). A candidate gene-based study in Douglas Fir revealed that 30 SNPs from 12 candidate genes out of 384 SNPs from 117 candidate genes were significantly associated with ten cold hardiness related traits (Eckert et al. 2009). Effect sizes of the significant SNPs were relatively small ranging from 1% to 3.6% of variance explained. In this study, the distributions of SNP effect sizes (percentage of the genetic variance explained by individual SNPs) highly concentrated near zero and few SNPs having large effects with maximum 28.8% explained genetic variation. A similar distribution of haplotype effect sizes was observed. However, validation of the effect sizes using an independent set of populations or near isogenic lines is needed to avoid bias. It is worth pointing out that only 170 SNPs in 12 candidate genes were used, a more complete picture of the genetic architecture of FT in rye might be seen using genome-wide association studies with high marker coverage.

4.2.3 Correlations of SNP-FT associations between platforms

Low to moderate empirical correlations of SNP-FT associations were observed across the three phenotyping platforms reflecting the complexity of FT. There are at least two reasons that might explain why relatively low to medium empirical correlations of SNP-FT associations were observed: 1) different duration and intensity of freezing temperature and 2) different levels of confounding effects from environmental factors other than frost stress *per se*. In the controlled platform, plants were cold-hardened and then exposed to freezing temperatures (-19°C or -21°C) in a short period of six days using defined temperature profiles.

Recovery score in the controlled platform represents the most pure and controlled measurement of FT among the three platforms since the effect of environmental factors other than frost stress is minimized. In the semi-controlled platform, plants were exposed to much longer freezing periods with fluctuating temperatures and repeated frost-thaw processes. In addition, a more complex situation occurs in this platform, requiring plants to cope with other variable climatic factors such as changing photoperiod, natural light intensity, wind, and limited water supply. Thus, the measurement % plants with undamaged leaves in the semi-controlled platform reflects the combined effect of various environmental influences and stresses on the vitality of leaf tissue. This measurement does not cover FT recover ability of the plant and is probably only suitable for mild frost stress. In the field platform, winter temperatures, similar to that in the semi-controlled platform, also fluctuated but were generally lower than in the semi-controlled platform due to strong continental climate in Eastern Europe and Canada. The measurement % survival in the field reflecting severe damage of the plant is further confounded by environmental effects, such as plant nutrition (Gusta et al. 1999), snow mould (Wisniewski et al. 1997), soil moisture (Szucs et al. 2003), topography, and other unmeasured factors. Phenotypic analysis in the field platform revealed significant effects of environment and genotype × environment, confirming the role of environmental factors on FT. It is worth pointing out that the variance of genotype × environment interaction was almost five times higher than the variance of genotype. Unfortunately, these environmental factors are difficult, if not impossible, to explicitly control in the field platform. One approach to control for environmental factors is to perform FT phenotyping in growth chambers where most of environmental effects are removed (e.g. wind) or can be monitored (e.g. temperature). Broad-sense heritability (h^2) in the controlled platform was considerably higher than in the field platform indicating better control of environmental factors. Another advantage of the controlled platform is that it allows more than one experiment per year and hence more than one selection cycle per year. Besides, the controlled platform guarantees sufficient frost stress, which is not always the case in field environments. However, given the low to moderate correlations among the three phenotyping platforms observed in both phenotypic and association analysis, predicting genotype performance only with semi-controlled and controlled platforms might be difficult since selection response is proportional to the correlation of FT between platforms (Lande and Thompson 1990). Further investigations, such as using the same endpoint, are needed to increase correlations among platforms.

4.2.4 Context-dependent effects of association analyses

Epistasis, generally defined as the interaction between genes, has been recognized for over a century (Bateson 1909), and recently it has been suggested that it should be explicitly modeled in association studies in order to detect "missing heritabilities" (Phillips 2008; Wu et al. 2010). Several recent association studies in plants have revealed the presence of epistasis in complex traits, including potato tuber quality, barley flowering time, and maize kernel quality (Li et al. 2010a; Manicacci et al. 2009; Stracke et al. 2009). In this study, eleven, six, and one significant ($P < 0.05$) gene × gene interaction effects were found in the controlled, semi-controlled and field platforms, respectively, suggesting that epistasis may play a role in the frost responsive network. From the frost responsive network, one might hypothesize that transcription factors interact with their downstream target genes, for example, that *ScIce2* interacts with the *ScCbf* gene family and the latter interacts with *COR* genes, such as the dehydrin (*Dhn*) gene family. Indeed significant interactions were observed between *ScIce2* × *ScCbf15*, *ScCbf14* × *ScDhn3*, and *ScDreb2* × *ScDhn3*. Some candidate genes in the same cascade level also interact with each other, such as members of the *ScCbf* gene family *ScCbf6* × *ScCbf15* and *ScCbf11* × *ScCbf14*. Similar interactions within the *Cbf* gene family were also observed in *Arabidopsis* where *AtCbf2* was indicated as a negative regulator of *AtCbf1* and *AtCbf3* (Novillo et al. 2004). In this study, *ScVrn1* was not significantly associated with FT but had significant interaction effects with six other candidate genes, suggesting an important role of *ScVrn1* in the frost responsive network. To confirm direct physical interactions of transcription factors with their downstream target genes, further experiments are needed, for example, electrophoresis mobility shift assays or chromatin immunoprecipitation (ChIP) sequencing technology. It is worth pointing out that the power of detecting gene × gene interaction in this study might have been low due to the relatively small sample size.

SNP or QTL by environment interaction is sometimes regarded as noise because the effects are not consistent across environments and only relevant in a specific environment. The presence of QTL by environment interaction depends on traits under study. In maize, detection of QTL for grain yield, plant height, and ear height highly depend on environment, whereas flowering-related traits such as anthesis-silking interval were less environmental dependent (Boer et al. 2007; Lima et al. 2006; Vargas et al. 2006). In this study, 54.7%, 38.8%, and 30.0% of the 170 SNPs showed significant SNP by environment interaction in the controlled, semi-controlled, and field platforms, respectively, indicating a need to consider SNP by environment interaction in MAS (Additional table 3). This high percentage of SNPs with significant SNP by environment interaction is not surprising since genotype by

environment interaction is highly significant ($P < 0.001$) in phenotypic analysis. Fifteen SNPs from three genes (*ScCbf9b*, *ScCbf12*, and *ScDreb2*) in the controlled platform, six SNPs from four genes (*ScCbf2*, *ScCbf12*, *ScCbf15* and *ScIce2*) in the semi-controlled platform, and three SNPs from three genes (*ScCbf9b*, *ScCbf12*, and *ScDhn1*) in the field platform were associated with FT with significant SNP main effect but without significant SNP by environment interaction effect implying that these SNP alleles consistently confer FT across environments. They are good marker candidates since the major goal of this study was to identify alleles associated with superior FT for selection of cultivars adapted to the wide range of climates in Eastern Europe. It is worth pointing out that among all 170 SNPs, SNP17 in *ScCbf12* was significantly associated with FT with SNP main effect across the three platforms and at the same time without significant SNP by environment interaction effect across the three platforms making it the best candidate for MAS.

4.2.5 Comparison between SNP- and haplotype-FT association analyses

Single-SNP association, namely testing one SNP at a time, has been proven to be powerful in the SNP-FT association analyses as shown in chapter 3.5.1. However, there is relatively little information provided by single SNPs unless the trait is monomorphic and the single SNP is the causal variant or highly correlated with it. For a complex quantitative trait like FT, simultaneous analyses of multiple SNPs may jointly provide information on associations leading to higher power. However, a statistical problem, the so-called "colinearity" may arise since predictors for the phenotype are highly correlated, e.g. high LD between SNPs in some of the candidate genes such as *ScCbf11*, *ScCbf14*, and *ScIce2*. An alternative method to perform multiple-SNP association analyses is to replace SNPs with haplotypes which are defined in this study as combinations of alleles from different SNPs within one gene. Haplotype-association has several advantages over SNP-association. Firstly, unlike the bi-allelic property of SNPs, haplotypes are often multi-allelic and therefore fit the distribution of quantitative traits better, thus increasing the statistical power. In other words, testing individual SNPs once at a time might neglect their joint distribution. Second, genetic variation in populations is structurally organized into haplotypes (Clark 2004). That is several SNPs might function together to change the protein structure if they are located in the coding region or capture the combined effect of *cis*-acting variants if they are located in the promoter region. Third, haplotype-association can circumvent the multiple testing problem by reducing the number of tests and thus increase the chance to reject the null hypothesis (Zhao et al. 2007). Fourth, haplotype-association can avoid colinearity in multiple SNP-associations. Due to these reasons, haplotype-association is widely used in human GWAS (Clark 2004; Morris

and Cardon 2007). Two challenges are often faced by the human geneticist: Defining haplotpye blocks with genome-wide distributed SNPs (Zhao et al. 2003) and inferring haplotype phase in heterozygous species (Clark 2004; Scheet and Stephens 2006; Stephens et al. 2001). In this study, it seems reasonable to define haplotpye blocks within the genes which are supposed to be the functional units in the genome. The employment of gamete capture gives rise to an unprecedented opportunity to determine haplotype phase in a heterozygous species without ambiguity. Despite many advantages of haplotype-association mentioned above, in this study less statistical power was observed in general compared to the SNP-association results except for *ScCbf2*. One possible reason is that there are many missing data ranging from 2.1% to 87.9% in haplotype-associations since the rare haplotypes (MAF < 0.05) were excluded from the analyses due to the possibility of causing type I error (false positive) inflation. Therefore, haplotype-associations in this study have a smaller sample size and consequently less statistical power (Table 10). A much larger sample size of germplasm is needed if one wants to test all rare haplotypes. Another reason could be that if the association signal is mainly driven by a single SNP, using haplotypes might just add noise to the association since LD in this case probably arose from genetic drift instead of selection (Hayes 2007). However, this explanation might be less likely because FT is a polygenic trait with many causal variants. Some other studies based on both simulation and real data suggested that single SNP-association had similar or greater power and precision than haplotype-association (Grapes et al. 2004; Nielsen et al. 2004; Zhao et al. 2007). However, Hayes (2007) came to an opposite conclusion and suggested that whether to use single SNP-association or haplotype-association is mainly determined by the level of LD in the population. If LD measured by r^2, is high or even equal to 1, testing single SNPs has a similar or the same result as haplotypes since they have a similar or even the same allele frequency; If LD is low, a different picture of association might arise. The present study supports this assumption that single SNP-association and haplotype-association in *ScIce2* with high LD had similar results whereas in *ScCbf2* with low LD more significant associations were found in haplotype-association than in single SNP-association. In conclusion, SNP and haplotype-association should both be performed whenever possible.

4.2.6 Multiple testing in association analyses

An $\alpha = 0.05$ has been widely accepted as the significance threshold in hypothesis testing (Sterne and Davey Smith 2001). Some studies also used $\alpha = 0.01$ as the significance threshold. Nevertheless, these significance thresholds are suitable if only one hypothesis is tested at a time. An inherent problem of hypothesis testing is the so-called “multiple testing”. Multiple

testing leads to an inflation of the false positive rate (type I error) which means a true null-hypothesis is rejected. A simple and common way to handle this problem is Bonferroni correction where the significance level is divided by the number of tests (Van Belle et al. 2004). However, the Bonferroni correction is very conservative and only suitable when independent factors are tested. This assumption is violated in this study due to high LD between SNPs in some genes such as *ScCbf11*, *ScCbf14*, and *ScIce2*. In order to retain SNP-candidates for further validation in upcoming experiments, the less stringent significance level of $\alpha = 0.05$ was taken as significance threshold. In Table 11 a summary on the numbers of significant associations at different alpha levels including Bonferroni correction is shown. Compared to the controlled platform, the numbers of significant associations in the semi-controlled and field platforms decreased faster when increasing the stringency of the significance level. In the controlled platform, nine SNPs from *ScCbf9b* and *ScCbf12*, two haplotypes from *ScCbf2*, and epistatic effects between *ScCbf6* × *ScCbf15*, *ScCbf6* × *ScVrn1*, and *ScCbf14* × *ScDhn3* remained significantly associated with FT even under the most stringent significance level ($\alpha = 0.000058$). Overall, the *P*-values of SNP and haplotype associations were smaller in the controlled platform compared to that in the semi-controlled and field platforms. One possible explanation is probably a smaller experimental error due to a better control of environmental noise in the controlled platforms. This is supported by the phenotypic data analyses where the heritability in the controlled platforms was almost twice as high as that in the field platform.

Table 11 Numbers of significant associations at different α levels

	Controlled			**Semi-controlled**			**Field**		
α level	SNPs (genes) [a]	Haplo-types	Epistasis	SNPs (genes)	Haplo-types	Epistasis	SNPs (genes)	Haplo-types	Epistasis
0.05	69 (8)	6	11	22 (4)	2	6	31 (6)	2	2
0.01	49 (6)	3	8	5 (3)	1	4	22 (3)	0	0
0.00416[b]	44 (6)	3	6	3 (3)	0	0	7 (2)	0	0
0.000294[c]	18 (4)	2	4	0	0	0	1(1)	0	0
0.000058[d]	9 (2)	2	3	0	0	0	0	0	0

[a] The numbers of genes were counted once there was a significant SNP present in the genes

[b] Adjusted for Bonferroni correction by 0.05/12 (the number of genes)

[c] Adjusted for Bonferroni correction by 0.05/170 (the number of SNPs) for SNP and by 0.05/30 for haplotype and epistasis

[d] Adjusted for Bonferroni correction by 0.01/170 for SNP and by 0.01/30 for haplotype and epistasis

4.3 Outlook

Fast and precise phenotyping of traits is crucial for association studies and other breeding experiments. Under field conditions, sufficiently low temperatures and durable frost periods are needed to obtain differentiation between frost tolerant and frost susceptible genotypes for association studies. Within six tested field environments, only two (SAS1 and SAS2) showed satisfactory FT differentiation between genotypes. This underlines the need of multiple years and locations for assessment for FT under field conditions. Besides, large experimental error due to inhomogeneity of soil and microclimate conditions poses a challenge in field trials. To address this, one approach is to use spatial models to adjust for the spatial trend across the field. As suggested by Gilmour et al. (1997), spatial variation can be partitioned into the following three additive components: local trend, global trend and extraneous variation. Another challenge of FT phenotyping is the need to perform precise and high-throughput phenotyping in order to catch up with the high-throughput genotyping technologies. For example, chlorophyll fluorescence which measures frost damage of leaves as a decrease in the maximum quantum efficiency of photosystem II could be automated using the high-throughput screening system, *Scanalyzer 3D*, provided by the LemnaTec company (Tester and Langridge 2010).

Genome-wide association studies, using high-density genotyping arrays to perform association analysis without prior knowledge, is an alternative to candidate gene-based association studies. This approach has successfully identified many QTL controlling morphological and agronomic traits in *Arabidopsis*, barley, and rice (Atwell et al. 2010; Cockram et al. 2011; Huang et al. 2010; Li et al. 2010b). For traits such as disease and insect resistance (Ingvarsson and Street 2010) that are regulated by a few genes with large effect sizes, it is logical to use single marker regression since the aim is to identify a few causal variants or markers in LD with the causal variants. In this approach markers are tested one at a time and then model selection techniques such as forward selection, backward elimination and/or stepwise regression are tested to select the best combination of markers. However, estimated marker effects are biased due to model selection. Therefore it is important to validate the significance and effect size of the markers in an independent population. Besides, the risk of identifying false positive signals due to multiple testing is high. For traits that are regulated by many genes with small effect sizes, fitting a multiple marker regression model using all markers simultaneously is an alternative to select a few significant markers. This method is known as genomic selection, a rapid evolving field of research (Bernardo and Yu 2007; Meuwissen et al. 2001). However analysis of such large amounts of markers together

leads to over-parameterisation where the number of observations is much smaller than the number of predictors (n<p). Various methods have been developed to tackle this problem, for example, ridge regression (Piepho 2009), Bayes A and Bayes B (Meuwissen et al. 2001), and Bayesian LASSO (Li et al. 2011). In addition, to select the right statistical model, another obvious concern for successful genome-wide association studies is the availability of high-density genotyping arrays generated in low cost, high-throughput, and high accuracy. Recently a Rye5K SNP array has been developed using rye transcriptome sequencing (Haseneyer et al. 2011) and a 20K SNP array will be available in the near future (Eva Bauer, pers. communication). It is exciting to see whether genome-wide association studies can help to identify new QTL controlling important agronomic traits in rye as well.

Classical plant breeding, relying solely on phenotypic selection has been historically successful since centuries. However, the disadvantages of phenotypic selection are obvious: Phenotypic selection is time-consuming, it is difficult to measure traits such as abiotic stress tolerance, and is dependent on the stage of development (e.g. yield can only be measured in the late stage of plant development). MAS, generally defined as a technique that utilizes DNA markers for selection of desirable genotypes, might overcome the disadvantages of phenotypic selection as mentioned above (Collard and Mackill 2008). It has been shown to be a very valuable tool for improving simply inherited traits such as pest or disease resistances (Jefferies et al. 2003) or even complex traits such as drought stress tolerance (Ribaut and Ragot 2007). MAS has the potential of providing more efficient use of new genetic variation from exotic germplasm. For example, aiming at broadening the genetic diversity of elite hybrid rye, an introgression library was developed by crossing an inbred line and a heterozygous Iranian primitive population followed by MAB using AFLP and SSR markers (Falke et al. 2008). SNPs and haplotypes with significant association with FT in the present study could be used as DNA markers for rapid and precise germplasm screening. Since most of the available genetic resources of rye are from population varieties, it will greatly facilitate the process of preselecting frost tolerant donor plants for hybrid rye breeding programs (Viktor Korzun, pers. communication). However, a validation step is needed before using these SNPs and haplotypes as DNA markers. Near-isogenic lines (NILs) which have a similar genetic background, but are homozygous for either the positive or negative alleles can be compared in the controlled and/or field platforms for FT. As an alternative validation approach, an allele-specific expression assay can be performed to assess differential allelic expression (Pastinen 2010; Serre et al. 2008). The relatively higher costs of MAS compared to traditional phenotypic selection is the main concern of Dreher et al. (2003), however, the

fast development of high-throughput genotyping platforms (e.g. Illumina's iScan) and next generation sequencing techniques (e.g. Roche's 454 GS FLX, Illumina's HiSeq2000, and ABI's SOLiD) plus the rapid decline of their costs will ultimately lead to a more common adoption of MAS in breeding programs in the near future.

5. Summary

Frost is an important abiotic stress that not only limits geographic distribution of crop production but also adversely affects crop development and yield. Crop varieties with improved frost tolerance (FT) are of enormous value for countries with severe winters. As the most frost tolerant small grain cereal, rye (*Secale cereale* L.) is an ideal cereal model for investigating the genetic basis of FT, a complex trait with polygenic inheritance. In order to dissect FT, a multi-platform candidate gene-based association approach was performed in 201 winter rye genotypes.

Plant material was derived from one Middle and four Eastern European cross-pollinated winter rye breeding populations: Petkus (Germany), PR (Belarus), EKO (Poland), SMH (Poland), and ROM (Poland). In order to determine the haplotype phase, gamete capture was performed where heterozygous plants from the five populations were crossed with the self-fertile inbred line Lo152 resulting in 201 heterozygous S_0 plants, each with one gamete known.

Since FT is a complex quantitative trait affected by many genetic and environmental factors, twelve environments from three different phenotyping platforms under controlled, semi-controlled, and field conditions were investigated. Significant genotypic variation of FT was found in all environments except for two field environments. However, the correlations of FT among the three different phenotyping platforms low to medium (r = 0.19-0.38) and thus might hinder the prediction of FT in the field from the other two platforms. Broad-sense heritability in the controlled platform was twice as high as that in the field platform indicating better control of environmental factors. Different populations might exhibit different FT due to local adaptation. The PR population showed slightly higher FT compared to other populations.

Twelve candidate genes with a putative role in the frost responsive network were studied including seven members of the *C-repeat Binding Factor* (*ScCbf*) transcription factor family, as well as *Dehydration-Responsive Element Binding gene 2*, (*ScDreb2*), dehydrin genes (*ScDhn1* and *ScDhn3*), *Inducer of Cbf Expression 2* (*ScIce2*), and vernalization gene *ScVrn1*. A total of 161 single nucleotide polymorphisms (SNPs) and 9 insertions and deletions (Indels) were found within around 10 kb of DNA sequence, resulting in an average polymorphism frequency of 1 polymorphism / 55 bp and an average nucleotide diversity $\pi = 5.6\times10^{-3}$. A high level of genetic variation within the germplasm was observed which was mainly attributed to within population variation. Using 37 genome-wide SSR markers, a reduced

level of genetic diversity in the Petkus population was observed. Population structure and kinship analyses further revealed that the Petkus population was distinguished from the Eastern European populations. Extent of linkage disequilibrium (LD) determines mapping resolution and marker density in association studies. The extent of LD in this germplasm over all genes and populations was very low with approximately 520 bp using $r^2 = 0.16$ as a critical threshold.

The major goal of this study was to identify and characterize favorable alleles conferring superior FT in winter rye using linear mixed models. Statistically significant ($P < 0.05$) associations between FT and SNPs or haplotypes of nine candidate genes were identified. Two SNPs in *ScCbf15* and one in *ScCbf12*, all leading to amino acid exchanges, were significantly associated with FT over all three phenotyping platforms. Favorable haplotypes in *ScIce2* and *ScCbf2* were mainly present in the PR population but entirely absent in the target Petkus population. Distribution of SNP effect sizes expressed as percentage of the genetic variance explained by individual SNPs was highly concentrated near zero with a few SNPs obtaining large effects. Relatively low to medium empirical correlations of SNP-FT associations were observed across the three platforms indicating the need for multi-level experimentation to dissect the complex mechanism of FT in rye. Two-way epistasis was found between 14 pairs of candidate genes suggesting the presence of epistatic interactions between genes involved in the frost responsive network. Significant ($P < 0.05$) SNP by environment interactions were found in more than 85% of the 170 SNPs in all candidate genes except for *ScVrn1*.

Identification of alleles and genes underlying agronomic traits is important for genome-based breeding. The results demonstrated that given the huge genome size of rye (~8,000 Mb) and the rapid decline of LD, the candidate gene-based association approach remains one of the most appropriate strategies for identification of alleles influencing agronomic traits.

6. Acknowledgements

First of all, I want to praise God for being by my side as I completed my PhD study. I would like to thank Prof. Dr. Chris-Carolin Schön for her guidance and critical examination of this work. I am grateful for having Prof. Donna P. Ankerst as my second supervisor as she gave me helpful advice and showed me the beauty of statistics. This project would not have been possible without the help of Drs. Eva Bauer and Grit Haseneyer (Chair of Plant Breeding), I very much appreciate their time and effort. I want to thank all my colleagues at the Chair of Plant Breeding for their continued support.

Interdisciplinary research cannot be done alone. No phenotyping, no association study! I am grateful to Dr. Peer Wilde (KWS LOCHOW Bergen, Germany), Prof. Dr. Gabor Galiba (ARI, Martonvásár, Hungary), Prof. Dr. Brian Fowler (University of Saskatoon, Saskatoon, Canada), Prof. Dr. Thomas Miedaner and Dr. Christof Kling (Universität Hohenheim, Stuttgart, Germany) for their participation in phenotyping. No statistics, no association study! Many thanks also to Andreas Böck (Biostatistics Unit, Technische Universität München, Freising, Germany) for giving me help in statistics whenever I needed it. No genotyping, no association study! Thanks to Dr. Viktor Korzun (KWS LOCHOW Bergen, Germany) for providing SSR data and thanks to Susanne Schrack and Tobias Dreser for technical assistance in the lab.

I would like to thank my parents and sister, who have always been and continue to be there for me in all my endeavors. And last but not least, I want to express my greatest gratitude to my wife, Man-tzu Hung, for her unwavering support during the writing of this thesis, even as our lovely daughter, Diana Li, came into this world. This thesis is dedicated to them both.

7. References

Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. Molecular Breeding 19:341-356

Alonso-Blanco C, Gomez-Mena C, Llorente F, Koornneef M, Salinas J, Martinez-Zapater JM (2005) Genetic and molecular analyses of natural variation indicate *CBF2* as a candidate gene for underlying a freezing tolerance quantitative trait locus in Arabidopsis. Plant Physiology 139:1304-1312

Aranzana MJ, Kim S, Zhao KY, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang CL, Toomajian C, Traw B, Zheng HG, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. Plos Genetics 1:531-539

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng DZ, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang CL, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627-631

Badawi M, Reddy YV, Agharbaoui Z, Tominaga Y, Danyluk J, Sarhan F, Houde M (2008) Structure and functional analysis of wheat *ICE* (inducer of *CBF* expression) genes. Plant and Cell Physiology 49:1237-1249

Baga M, Chodaparambil SV, Limin AE, Pecar M, Fowler DB, Chibbar RN (2007) Identification of quantitative trait loci and associated candidate genes for low-temperature tolerance in cold-hardy winter wheat. Functional & Integrative Genomics 7:53-68

Bao JS, Corke H, Sun M (2006) Microsatellites, single nucleotide polymorphisms and a sequence tagged site in starch-synthesizing genes in relation to starch physicochemical properties in nonwaxy rice (*Oryza sativa* L.). Theoretical and Applied Genetics 113:1185-1196

Bates D, Maechler M (2010) lme4: Linear mixed-effects models using S4 classes. R package version 0999375-35

Bateson W (1909) Mendel's principles of heredity. Cambridge University Press, Cambridge

Beales J, Laurie DA, Devos KM (2005) Allelic variation at the linked *AP1* and *PhyC* loci in hexaploid wheat is associated but not perfectly correlated with vernalization response. Theoretical and Applied Genetics 110:1099-1107

Bernardo R, Yu JM (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Science 47:1082-1090

Boer MP, Wright D, Feng LZ, Podlich DW, Luo L, Cooper M, van Eeuwijk FA (2007) A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics 177:1801-1813

Bolaric S, Barth S, Melchinger AE, Posselt UK (2005) Genetic diversity in European perennial ryegrass cultivars investigated with RAPD markers. Plant Breeding 124:161-166

Bolibok-Bragoszewska H, Heller-Uszynska K, Wenzl P, Uszynski G, Kilian A, Rakoczy-Trojanowska M (2009) DArT markers for the rye genome - genetic diversity and mapping. BMC Genomics 10:11

Bondia-Pons I, Aura AM, Vuorela S, Kolehmainen M, Mykkanen H, Poutanen K (2009) Rye phenolics in nutrition and health. Journal of Cereal Science 49:323-336

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32:314-331

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-1177

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 HH, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Villeda HS, da Silva HS, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu JM, Zhang ZW, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science 325:714-718

Bushuk W (2001) Rye: Production, chemistry, and technology, 2nd edn. American Association of Cereal Chemists Inc., St. Paul, MN, USA

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

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

Campoli C, Matus-Cadiz MA, Pozniak CJ, Cattivelli L, Fowler DB (2009) Comparative expression of *Cbf* genes in the *Triticeae* under different acclimation induction temperatures. Mol Genet Genomics 282:141-152

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 Genetics 3:19

Chinnusamy V, Zhu J, Zhu JK (2007) Cold stress regulation of gene expression in plants. Trends in Plant Science 12:444-451

Choi DW, Zhu B, Close TJ (1999) The barley (*Hordeum vulgare L*.) dehydrin multigene family: sequences, allele types, chromosome assignments, and expression characteristics of 11 *Dhn* genes of cv Dicktoo. Theoretical and Applied Genetics 98:1234-1247

Clark AG (2004) The role of haplotypes in candidate gene studies. Genetic Epidemiology 27:321-333

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo ZW, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WTB, Ramsay L, Mackay I, Balding DJ, Waugh R, O'Sullivan DM (2011) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proceedings of the National Academy of Sciences of the United States of America 107:21611-21616

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B-Biological Sciences 363:557-572

Cook D, Fowler S, Fiehn O, Thomashow MF (2004) A prominent role for the *CBF* cold response pathway in configuring the low-temperature metabolome of *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America 101:15243-15248

Cuesta-Marcos A, Szucs P, Close TJ, Filichkin T, Muehlbauer GJ, Smith KP, Hayes PM (2010) Genome-wide SNPs and re-sequencing of growth habit and inflorescence genes

in barley: implications for association mapping in germplasm arrays varying in size and structure. BMC Genomics 11:707

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997-1004

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

Dmitriev DA, Rakitov RA (2008) Decoding of superimposed traces produced by direct sequencing of heterozygous Indels. Plos Computational Biology 4:7

Dorffling K, Schulenburg S, Lesselich G, Dorffling H (1990) Abscisic acid and proline levels in cold hardened winter wheat leaves in relation to variety-specific differences in freezing resistance. Journal of Agronomy and Crop Science 165:230-239

Dreher K, Khairallah M, Ribaut JM, Morris M (2003) Money matters (I): costs of field and laboratory procedures associated with conventional and marker-assisted maize breeding at CIMMYT. Molecular Breeding 11:221-234

Dubouzet JG, Sakuma Y, Ito Y, Kasuga M, Dubouzet EG, Miura S, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) *OsDREB* genes in rice, *Oryza sativa* L., encode transcription activators that function in drought-, high-salt-, and cold-responsive gene expression. Plant Journal 33:751-763

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, Clair JBS, Neale DB (2009) Asssociation genetics of coastal douglas fir (Pseudotsuga menziesu var. Menziesii, Pinaceae). I. Cold-hardiness related traits. Genetics 182:1289-1302

Egawa C, Kobayashi F, Ishibashi M, Nakamura T, Nakamura C, Takumi S (2006) Differential regulation of transcript accumulation and alternative splicing of a *DREB2* homolog under abiotic stress conditions in common wheat. Genes & Genetic Systems 81:77-91

Ehrenreich IM, Hanzawa Y, Chou L, Roe JL, Kover PX, Purugganan MD (2009) Candidate gene association mapping of *Arabidopsis* flowering time. Genetics 183:325-335

Ehrenreich IM, Stafford PA, Purugganan MD (2007) The genetic architecture of shoot branching in Arabidopsis thaliana: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping. Genetics 176:1223-1236

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics 1:47-50

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491

Falke KC, Susic Z, Hackauf B, Korzun V, Schondelmaier J, Wilde P, Wehling P, Wortmann H, Mank R, van der Voort JR, Maurer HP, Miedaner T, Geiger HH (2008) Establishment of introgression libraries in hybrid rye (*Secale cereale* L.) from an Iranian primitive accession as a new tool for rye breeding and genomics. Theoretical and Applied Genetics 117:641-652

Figueiredo LFD, Sine B, Chantereau J, Mestres C, Fliedel G, Rami JF, Glaszmann JC, Deu M, Courtois B (2010) Variability of grain quality in sorghum: association with polymorphism in *Sh2, Bt2, SssI, Ae1, Wx* and *O2*. Theoretical and Applied Genetics 121:1171-1185

Fischer S, Melchinger AE, Korzun V, Wilde P, Schmiedchen B, Möhring J, Piepho HP, Dhillon BS, Würschum T, Reif JC (2010) Molecular marker assisted broadening of the Central European heterotic groups in rye with Eastern European germplasm. Theoretical and Applied Genetics 120:291-299

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

Flint J, Mackay TFC (2009) Genetic architecture of quantitative traits in mice, flies, and humans. Genome Research 19:723-733

Fowler DB, Gusta LV, Tyler NJ (1981) Selection for winterhardiness in wheat. III. Screening methods. Crop Science 21:896-901

Fowler DB, Limin AE (1987) Exploitable genetic variability for cold tolerance in commercially grown cereals. Canadian Journal of Plant Science 67:278-278

Fowler S, Thomashow MF (2002) *Arabidopsis* transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the *CBF* cold response pathway. Plant Cell 14:1675-1690

Francia E, Barabaschi D, Tondelli A, Laido G, Rizza F, Stanca AM, Busconi M, Fogher C, Stockinger EJ, Pecchioni N (2007) Fine mapping of a *HvCBF* gene cluster at the frost resistance locus *Fr-H2* in barley. Theoretical and Applied Genetics 115:1083-1091

Fricano A, Rizza F, Faccioli P, Pagani D, Pavan P, Stella A, Rossini L, Piffanelli P, Cattivelli L (2009) Genetic variants of *HvCbf14* are statistically associated with frost tolerance in a European germplasm collection of *Hordeum vulgare*. Theoretical and Applied Genetics 119:1335-1348

Fursova OV, Pogorelko GV, Tarasov VA (2009) Identification of *ICE2*, a gene involved in cold acclimation which determines freezing tolerance in *Arabidopsis thaliana*. Gene 429:98-103

Galiba G, Quarrie SA, Sutka J, Morgounov A, Snape JW (1995) RFLP mapping of the vernalization (*Vrn1*) and frost resistance (*Fr1*) genes on chromosome 5A of wheat. Theoretical and Applied Genetics 90:1174-1179

Galiba G, Vagujfalvi A, Li CX, Soltesz A, Dubcovsky J (2009) Regulatory genes involved in the determination of frost tolerance in temperate cereals. Plant Science 176:12-19

George J, Dobrowolski MP, de Jong EV, Cogan NOI, Smith KF, Forster JW (2006) Assessment of genetic diversity in cultivars of white clover (*Trifolium repens* L.) detected by SSR polymorphisms. Genome 49:919-930

Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. Journal of Agricultural, Biological, and Environmental Statistics 2:269-293

Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics. Journal of Heredity 86:485-486

Grapes L, Dekkers JCM, Rothschild MF, Fernando RL (2004) Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. Genetics 166:1561-1570

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

Gusta LV, O'Connor BJ, Lafond GL (1999) Phosphorus and nitrogen effects on the freezing tolerance of Norstar winter wheat. Canadian Journal of Plant Science 79:191-195

Gusta LV, Willen R, Fu P, Robertson AJ, Wu GH (1997) Genetic and environmental control of winter survival of winter cereals. Acta Agronomica Hungarica 45:231-240

Gustafson JP, Ma XF, Korzun V, Snape JW (2009) A consensus map of rye integrating mapping data from five mapping populations. Theoretical and Applied Genetics 118:793-800

Hackauf B, Wehling P (2002) Identification of microsatellite polymorphisms in an expressed portion of the rye genome. Plant Breeding 121:17-25

Hamblin MT, Fernandez MGS, 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

Hannah MA, Heyer AG, Hincha DK (2005) A global survey of gene regulation during cold acclimation in *Arabidopsis thaliana*. Plos Genetics 1:179-196

Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, Stapleton AE, Vallabhaneni R, Williams M, Wurtzel ET, Yan JB, Buckler ES (2008) Natural genetic

variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319:330-333

Haseneyer G, Schmutzer T, Seidel M, Zhou R, Mascher M, Schon CC, Taudien S, Scholz U, Stein N, Mayer KF, Bauer E (2011) From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). BMC Plant Biology 11:131

Haseneyer G, Stracke S, Piepho HP, Sauer S, Geiger HH, Graner A (2010) DNA polymorphisms and haplotype patterns of transcription factors involved in barley endosperm development are associated with key agronomic traits. BMC Plant Biology 10:11

Hayes B (2007) QTL mapping, MAS, and genomic selection. Department of primary industries, Victoria, pp 43-44

Hayes BJ, Goddard ME (2008) Technical note: Prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science 86:2089-2092

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38:226-231

Hill WG, Weir BS (1988) Variances and covariance of squared linkage disequilibrium in finite populations. Theoretical Population Biology 33:54-78

Hillman G, Hedges R, Moore A, Colledge S, Pettitt P (2001) New evidence of Lateglacial cereal cultivation at Abu Hureyra on the Euphrates. Holocene 11:383-393

Hömmö LM (1994) Hardening of some winter wheat (*Triticum aestivum* L.), rye (*Secale cereals* L.), triticale (*Triticosecale Wittmack*) and winter barley (*Hordeum vulgare* L.) cultivars during autumn and the final winter survival in Finland. Plant Breeding 112:285-293

Houde M, Dhindsa RS, Sarhan F (1992) A molecular marker to select for freezing tolerance in *Gramineae*. Molecular and General Genetics 234:43-48

Huang XH, Wei XH, Sang T, Zhao QA, Feng Q, Zhao Y, Li CY, Zhu CR, Lu TT, Zhang ZW, Li M, Fan DL, Guo YL, Wang A, Wang L, Deng LW, Li WJ, Lu YQ, Weng QJ, Liu KY, Huang T, Zhou TY, Jing YF, Li W, Lin Z, Buckler ES, Qian QA, Zhang QF, Li JY, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics 42:961-U976

Huld TA, Šúri M, Dunlop ED, Micale F (2006) Estimating average daytime and daily temperature profiles within Europe. Environmental Modelling & Software 21:1650-1661

Ingvarsson PK, Street NR (2010) Association genetics of complex traits in plants. New Phytologist 189:909-922

Jaglo KR, Kleff S, Amundsen KL, Zhang X, Haake V, Zhang JZ, Deits T, Thomashow MF (2001) Components of the *Arabidopsis* C-repeat/dehydration-responsive element binding factor cold-response pathway are conserved in *Brassica napus* and other plant species. Plant Physiology 127:910-917

Jefferies SP, King BJ, Barr AR, Warner P, Logue SJ, Langridge P (2003) Marker-assisted backcross introgression of the *Yd2* gene conferring resistance to barley yellow dwarf virus in barley. Plant Breeding 122:52-56

Jones N, Ougham H, Thomas H, Pasakinskiene I (2009) Markers and mapping revisited: finding your gene. New Phytologist 183:935-966

Jun TH, Van K, Kim MY, Lee SH, Walker DR (2008) Association analysis using SSR markers to find QTL for seed protein content in soybean. Euphytica 162:179-191

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723

Kaplan F, Kopka J, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of *Arabidopsis*. Plant Physiology 136:4159-4168

Khlestkina EK, Ma HMT, Pestsova EG, Roder MS, Malyshev SV, Korzun V, Börner A (2004) Mapping of 99 new microsatellite-derived loci in rye (*Secale cereale* L.) including 39 expressed sequence tags. Theoretical and Applied Genetics 109:725-732

Khush GS (1963) Cytogenetic and evolutionary studies in *Secale* III. Cytogenetics of weedy ryes and origin of cultivated rye. Economic Botany 17:60-71

Kilian B, Ozkan H, Kohl J, von Haeseler A, Barale F, Deusch O, Brandolini A, Yucel C, Martin W, Salamini F (2006) Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. Molecular Genetics and Genomics 276:230-241

Knox AK, Dhillon T, Cheng HM, Tondelli A, Pecchioni N, Stockinger EJ (2010) *CBF* gene copy number variation at Frost Resistance-2 is associated with levels of freezing tolerance in temperate-climate cereals. Theoretical and Applied Genetics 121:21-35

Knox AK, Li CX, Vagujfalvi A, Galiba G, Stockinger EJ, Dubcovsky J (2008) Identification of candidate *CBF* genes for the frost tolerance locus *Fr-A(m)2* in *Triticum monococcum*. Plant Molecular Biology 67:257-270

Kosova K, Prasil IT, Vitamvas P (2008) The relationship between vernalization-and photoperiodically-regulated genes and the development of frost tolerance in wheat and barley. Biologia Plantarum 52:601-615

Kosova K, Vitamvas P, Prasil IT (2007) The role of dehydrins in plant response to cold. Biologia Plantarum 51:601-617

Kreps JA, Wu YJ, Chang HS, Zhu T, Wang X, Harper JF (2002) Transcriptome changes for *Arabidopsis* in response to salt, osmotic, and cold stress. Plant Physiology 130:2129-2141

Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics 171:2029-2041

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756

Li J, Das K, Fu G, Li R, Wu R (2011) The Bayesian lasso for genome-wide association studies. Bioinformatics 27:516-523

Li L, Paulo MJ, Strahwald J, Lubeck J, Hofferbert HR, Tacke E, Junghans H, Wunder J, Draffehn A, van Eeuwijk F, Gebhardt C (2008) Natural DNA variation at candidate loci is associated with potato chip color, tuber starch content, yield and starch yield. Theoretical and Applied Genetics 116:1167-1181

Li L, Paulo MJ, van Eeuwijk F, Gebhardt C (2010a) Statistical epistasis between candidate gene alleles for complex tuber traits in an association mapping population of tetraploid potato. Theoretical and Applied Genetics 121:1303-1310

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO (2010b) Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America 107:21199-21204

Lima MDA, de Souza CL, Bento DAV, de Souza AP, Carlini-Garcia LA (2006) Mapping QTL for grain yield and plant traits in a tropical maize population. Molecular Breeding 17:227-239

Lin YH, Hwang SY, Hsu PY, Chiang YC, Huang CL, Wang CN, Lin TP (2008) Molecular population genetics and gene expression analysis of duplicated *CBF* genes of *Arabidopsis thaliana*. BMC Plant Biology 8:16

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128-2129

Liu Q, Kasuga M, Sakuma Y, Abe H, Miura S, Yamaguchi-Shinozaki K, Shinozaki K (1998) Two transcription factors, *DREB1* and *DREB2*, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-

temperature-responsive gene expression, respectively, in *Arabidopsis*. Plant Cell 10:1391-1406

Maccaferri M, Sanguineti MC, Noli E, Tuberosa R (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Molecular Breeding 15:271-289

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175:879-889

Manicacci D, Camus-Kulandaivelu L, Fourmann M, Arar C, Barrault S, Rousselet A, Feminias N, Consoli L, Frances L, Mechin V, Murigneux A, Prioul JL, Charcosset A, Damerval C (2009) Epistatic interactions between *Opaque2* transcriptional activator and its target gene *CyPPDK1* control kernel trait variation in maize. Plant Physiology 150:506-520

Mathews KL, Malosetti M, Chapman S, McIntyre L, Reynolds M, Shorter R, van Eeuwijk F (2008) Multi-environment QTL mixed models for drought stress adaptation in wheat. Theoretical and Applied Genetics 117:1077-1091

McKhann HI, Gery C, Berard A, Leveque S, Zuther E, Hincha DK, De Mita S, Brunel D, Teoule E (2008) Natural variation in *CBF* gene sequence, gene expression and freezing tolerance in the Versailles core collection of *Arabidopsis thaliana*. BMC Plant Biology 8:105

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829

Monroy AF, Dryanova A, Malette B, Oren DH, Farajalla MR, Liu W, Danyluk J, Ubayasena LWC, Kane K, Scoles GJ, Sarhan F, Gulick PJ (2007) Regulatory gene candidates and gene expression analysis of cold acclimation in winter and spring wheat. Plant Molecular Biology 64:409-423

Morris AP, Cardon LR (2007) Whole genome association. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of Statistical Genetics, 3rd edn. John Wiley & Sons, Ltd, West Sussex

Murphy D (2007) People, plants, and genes: the story of crops and humanity. Oxford University Press UK

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, NY

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

Novillo F, Alonso JM, Ecker JR, Salinas J (2004) *CBF2/DREB1C* is a negative regulator of *CBF1/DREB1B* and *CBF3/DREB1A* expression and plays a central role in, stress tolerance in *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America 101:3985-3990

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD (2004) Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics 167:1361-1369

Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. Nature Reviews Genetics 11:533-538

Persson K, Diaz O, von Bothmer R (2001) Extent and patterns of RAPD variation in landraces and cultivars of rye (*Secale cereale* L.) from Northern Europe. Hereditas 134:237-243

Persson K, von Bothmer R (2000) Assessing the allozyme variation in cultivars and Swedish landraces of rye (*Secale cereale* L.). Hereditas 132:7-17

Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. Conservation Biology 12:844-855

Phillips PC (2008) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9:855-867

Piepho HP (2009) Ridge Regression and Extensions for Genomewide Selection in Maize. Crop Science 49:1165-1176

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

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

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000c) Association mapping in structured populations. American Journal of Human Genetics 67:170-181

Qin F, Kakimoto M, Sakuma Y, Maruyama K, Osakabe Y, Tran LSP, Shinozaki K, Yamaguchi-Shinozaki K (2007) Regulation and functional analysis of *ZmDREB2A* in response to drought and heat stresses in *Zea mays L*. Plant Journal 50:54-69

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology 5:94-100

Reif JC, Gowda M, Maurer HP, Longin CFH, Korzun V, Ebmeyer E, Bothe R, Pietsch C, Würschum T (2011) Association mapping for quality traits in soft winter wheat. Theoretical and Applied Genetics 122:961-970

Revilla P, Butrón A, Cartea ME, Malvar RA, Ordás A (2005) Breeding for cold tolerance. In: Ashraf M, Harris PJC (eds) Abiotic stress: plant resistance through breeding and molecular approaches. Food Products Press, London, Oxford

Ribaut JM, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. Journal of Experimental Botany 58:351-360

Rogers JS (1972) Measures of genetic similarity and genetic distance. University of Texas Publication, Austin, Texas

Rogowsky PM, Guidet FLY, Langridge P, Shepherd KW, Koebner RMD (1991) Isolation and characterisation of wheat-rye recombinants involving chromosome arm 1DS of wheat. Theoretical and Applied Genetics 82:537-544

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497

Saidou AA, Mariac C, Luong V, Pham JL, Bezancon G, Vigouroux Y (2009) Association studies identify natural variation at *PHYC* linked to flowering time and morphological variation in pearl millet. Genetics 182:899-910

Salamini F, Ozkan H, Brandolini A, Schafer-Pregl R, Martin W (2002) Genetics and geography of wild cereal domestication in the Near East. Nature Reviews Genetics 3:429-441

Saulescu NN, Braun HJ (2001) Cold tolerance. In: Reynolds MP, Ortiz-Monasterio JI, McNab A (eds) Application of physiology in wheat breeding. CIMMYT, Mexico, D.F, pp 111-123

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78:629-644

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan JB, Hudson TJ (2008) Differential allelic expression in the human genome: A robust approach to identify genetic and epigenetic *Cis*-acting mechanisms regulating gene expression. Plos Genetics 4:(2): e1000006. doi:1000010.1001371/journal.pgen.1000006

Shang HY, Wei YM, Wang XR, Zheng YL (2006) Genetic diversity and phylogenetic relationships in the rye genus *Secale L*. (rye) based on *Secale cereale* microsatellite markers. Genetics and Molecular Biology 29:685-691

Shendure J, Ji HL (2008) Next-generation DNA sequencing. Nature Biotechnology 26:1135-1145

Skinner JS, von Zitzewitz J, Szucs P, Marquez-Cedillo L, Filichkin T, Amundsen K, Stockinger EJ, Thomashow MF, Chen THH, Hayes PM (2005) Structural, functional, and phylogenetic characterization of a large *CBF* gene family in barley. Plant Molecular Biology 59:533-551

Skot L, Humphreys MO, Armstead I, Heywood S, Skot KP, Sanderson R, Thomas ID, Chorlton KH, Hamilton NRS (2005) An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). Molecular Breeding 15:233-245

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68:978-989

Sterne JA, Davey Smith G (2001) Sifting the evidence-what's wrong with significance tests? British Medical Journal 322:226-231

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, Piepho HP, Schulz B, Melchinger AE (2008b) Multi-trait association mapping in sugar beet (*Beta vulgaris* L.). Theoretical and Applied Genetics 117:947-954

Stockinger EJ, Skinner JS, Gardner KG, Francia E, Pecchioni N (2007) Expression levels of barley *Cbf* genes at the Frost resistance-H2 locus are dependent upon alleles at *Fr-H1* and *Fr-H2*. Plant Journal 51:308-321

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. Theoretical and Applied Genetics 118:259-273

Sutton F, Chen DG, Ge XJ, Kenefick D (2009) *Cbf* genes of the *Fr-A2* allele are differentially regulated between long-term cold acclimated crown tissue of freeze-resistant and - susceptible, winter wheat mutant lines. BMC Plant Biology 9:34

Szucs P, Veisz O, Bedo Z (2003) Frost tolerance of *Triticum durum* (Desf.) genotypes in soils with various moisture contents. Cereal Research Communications 31:153-160

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays ssp mays* L.). Proceedings of the National Academy of Sciences of the United States of America 98:9161-9166

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818-822

Thomashow MF (1999) Plant cold acclimation: Freezing tolerance genes and regulatory mechanisms. Annual Review of Plant Physiology and Plant Molecular Biology 50:571-599

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

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics 43:159-U113

Tian F, Stevens NM, Buckler ES (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proceedings of the National Academy of Sciences of the United States of America 106:9979-9986

Vagujfalvi A, Galiba G, Cattivelli L, Dubcovsky J (2003) The cold-regulated transcriptional activator *Cbf3* is linked to the frost-tolerance locus *Fr-A2* on wheat chromosome 5A. Molecular Genetics and Genomics 269:60-67

Van Belle G, Fisher L, Heagerty P, Lumley T (2004) Biostatistics: A methodology for the health sciences, Second edn. John Wiley & Sons, Inc, Hoboken, New Jersey

Vargas M, van Eeuwijk FA, Crossa J, Ribaut JM (2006) Mapping QTLs and QTL x environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. Theoretical and Applied Genetics 112:1009-1023

Varshney RK, Beier U, Khlestkina EK, Kota R, Korzun V, Graner A, Börner A (2007) Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency, and applications for genome mapping and diversity studies. Theoretical and Applied Genetics 114:1105-1116

Vogel JT, Zarka DG, Van Buskirk HA, Fowler SG, Thomashow MF (2005) Roles of the *CBF2* and *ZAT12* transcription factors in configuring the low temperature transcriptome of *Arabidopsis*. Plant Journal 41:195-211

Wachowiak W, Balk PA, Savolainen O (2009) Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris L.*). Tree Genetics and Genomics 5:117-132

Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54:426-482

Wang J, McClean PE, Lee R, Goos RJ, Helms T (2008) Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. *Merr.*) advanced breeding lines. Theoretical and Applied Genetics 116:777-787

Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM, Buckler ESt (2004) Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719-2733

Winfield MO, Lu CG, Wilson ID, Coghill JA, Edwards KJ (2010) Plant responses to cold: transcriptome analysis of wheat. Plant Biotechnol J 8:749-771

Wisniewski K, Zagdanska B, Pronczuk M (1997) Interrelationship between frost tolerance, drought and resistance to snow mould (*Microdochium nivale*). Acta Agronomica Hungarica 45:311-316

Wu XS, Dong H, Luo L, Zhu Y, Peng G, Reveille JD, Xiong MM (2010) A novel statistic for genome-wide interaction analysis. Plos Genetics 6:(9)

Xing Y, Frei U, Schejbel B, Asp T, Lübberstedt T (2007) Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*. BMC Plant Biology 7:43

Xue GP, Loveridge CW (2004) *HvDRF1* is involved in abscisic acid-mediated gene regulation in barley and produces two forms of AP2 transcriptional activators, interacting preferably with a CT-rich element. Plant Journal 37:326-339

Yamaguchi-Shinozaki K, Shinozaki K (2006) Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. Annual Review of Plant Biology 57:781-803

Yamasaki M, Schroeder S, Sanchez-Villeda H, Gaut B, McMullen MD (2008) Empirical analysis of selection screens for domestication and improvement loci in maize by extended DNA sequencing. Plant Genome 1:33-43

Yan CJ, Tian ZX, Fang YW, Yang YC, Li JA, Zeng SY, Gu SL, Xu CW, Tang SZ, Gu MH (2011) Genetic analysis of starch paste viscosity parameters in glutinous rice (*Oryza sativa* L.). Theoretical and Applied Genetics 122:63-76

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proceedings of the National Academy of Sciences of the United States of America 100:6263-6268

Yu JM, Buckler ES (2006) Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17:155-160

Yu JM, 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. Nature Genetics 38:203-208

Zhang DL, Zhang HL, Wang MX, Sun JL, Qi YW, Wang FM, Wei XH, Han LZ, Wang XK, Li ZC (2009) Genetic structure and differentiation of *Oryza sativa L*. in China revealed by microsatellites. Theoretical and Applied Genetics 119:1105-1117

Zhao HH, Fernando RL, Dekkers JCM (2007) Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. Genetics 175:1975-1986

Zhao HY, Pfeiffer R, Gail MH (2003) Haplotype analysis in population genetics and association studies. Pharmacogenomics 4:171-178

Zhu B, Choi DW, Fenton R, Close TJ (2000) Expression of the barley dehydrin multigene family and the development of freezing tolerance. Molecular and General Genetics 264:145-153

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

Zhu JH, Shi HZ, Lee BH, Damsz B, Cheng S, Stirm V, Zhu JK, Hasegawa PM, Bressan RA (2004) An *Arabidopsis* homeodomain transcription factor gene, *HOS9*, mediates cold tolerance through a *CBF*-independent pathway. Proceedings of the National Academy of Sciences of the United States of America 101:9873-9878

Zonneveld BJM, Leitch IJ, Bennett MD (2005) First nuclear DNA amounts in more than 300 angiosperms. Annals of Botany 96:229-244

8. Additional tables and figures

Additional table 1 Genetic diversities of twelve candidate genes within five rye populations

	PR	EKO	SMH	ROM	Petkus
ScCbf2					
No. of genotypes	27	30	14	34	61
No. of polymorphisms	2	3	3	3	3
No. of haplotypes (private)	3 (0)	6 (0)	5 (0)	4 (0)	4 (1)
Hd± SD	0.62 ± 0.07	0.64 ± 0.05	0.78 ± 0.09	0.61 ± 0.06	0.70 ± 0.03
π ± SD ($\times 10^{-3}$)	1.4 ± 0. 2	1.3 ± 0.2	2.1 ± 0.4	1.2 ± 0.2	1.6 ± 0.1
ScCbf6					
No. of genotypes	32	42	15	36	69
No. of polymorphisms	3	5	3	5	3
No. of haplotypes (private)	3 (0)	7 (2)	2 (0)	6 (2)	4 (0)
Hd± SD	0.49 ± 0.09	0.62 ± 0.05	0.13 ± 0.11	0.31 ± 0.10	0.40 ± 0.06
π ± SD ($\times 10^{-3}$)	3.9 ± 0.6	5.0 ± 0.3	1.2 ± 0.1	2.1 ± 0.8	3.4 ± 0.5
ScCbf9b					
No. of genotypes	29	38	14	39	59
No. of polymorphisms	25	31	27	32	30
No. of haplotypes (private)	23 (12)	27 (17)	11 (5)	23 (15)	33 (26)
Hd± SD	0.98 ± 0.01	0.98 ± 0.01	0.93 ± 0.06	0.96 ± 0.02	0.96 ± 0.01
π ± SD ($\times 10^{-3}$)	6.1 ± 0.6	6.6 ± 0.6	6.8 ± 0.7	6.8 ± 0.6	6.9 ± 0.6
ScCbf11					
No. of genotypes	12	30	4	25	54
No. of polymorphisms	15	28	13	28	28
No. of haplotypes (private)	3 (1)	4 (1)	2 (0)	6 (3)	7 (4)
Hd± SD	0.32 ± 0.16	0.53 ± 0.09	0.50 ± 0.27	0.73 ± 0.06	0.60 ± 0.05
π ± SD ($\times 10^{-3}$)	4.0 ± 2.8	11.5 ± 2.6	10.4 ± 5.6	16.6 ± 2.2	12.7 ± 1.5
ScCbf12					
No. of genotypes	20	32	12	33	43
No. of polymorphisms	25	23	22	24	25
No. of haplotypes (private)	12 (8)	11 (4)	7 (3)	14 (8)	21 (15)
Hd± SD	0.92 ± 0.04	0.79 ± 0.05	0.83 ± 0.1	0.89 ± 0.04	0.91 ± 0.03
π ± SD ($\times 10^{-3}$)	11.7 ± 2.7	4.8 ± 2.0	6.3 ± 2.9	5.4 ± 1.3	12.7 ± 1.5
ScCbf14					
No. of genotypes	23	39	14	40	66
No. of polymorphisms	5	5	5	5	5
No. of haplotypes (private)	2 (0)	3 (1)	2 (0)	2 (0)	3 (1)
Hd± SD	0.09 ± 0.08	0.19 ± 0.08	0.14 ± 0.12	0.10 ± 0.06	0.24 ± 0.06
π ± SD ($\times 10^{-3}$)	0.8 ± 0.7	1.6 ± 0.7	1.3 ± 11	0.9 ± 0.6	2.1 ± 0.5

(Additional table 1 continued)

	PR	EKO	SMH	ROM	Petkus
ScCbf15					
No. of genotypes	28	41	13	37	49
No. of polymorphisms	4	2	2	4	4
No. of haplotypes (private)	7 (2)	3 (0)	3 (0)	4 (0)	6 (2)
Hd± SD	0.83 ± 0.04	0.26 ± 0.08	0.50 ± 0.14	0.70 ± 0.04	0.69 ± 0.04
π ± SD ($\times 10^{-3}$)	3.5 ± 0.3	0.5 ± 0.3	1.1 ± 0.3	3.4± 0.2	3.3 ± 0.2
ScDhn1					
No. of genotypes	18	35	11	28	44
No. of polymorphisms	5	6	0	6	6
No. of haplotypes (private)	4 (1)	7 (2)	1 (0)	6 (3)	10 (7)
Hd± SD	0.48 ± 0.13	0.56 ± 0.09	0	0.39 ± 0.12	0.41 ± 0.09
π ± SD ($\times 10^{-3}$)	10.8 ± 4.4	6.0 ± 1.7	0	2.2 ± 1.0	5.8 ± 2.0
ScDhn3					
No. of genotypes	23	23	13	21	49
No. of polymorphisms	12	7	10	8	4
No. of haplotypes (private)	10 (6)	8 (1)	10 (6)	5 (1)	7 (4)
Hd ± SD	0.84 ± 0.06	0.77 ± 0.06	0.95 ± 0.05	0.61 ± 0.09	0.66 ± 0.04
π ± SD ($\times 10^{-3}$)	14.2 ± 1.6	7.5 ± 1.6	14.1 ± 2.0	10.2 ± 2.5	5.8 ± 0.5
ScDreb2					
No. of genotypes	18	41	15	38	64
No. of polymorphisms	8	5	6	7	7
No. of haplotypes (private)	18 (12)	17 (12)	12 (6)	22 (15)	34 (28)
Hd ± SD	0.84 ± 0.06	0.81 ± 0.00	0.96 ± 0.00	0.92 ± 0.00	0.92 ± 0.00
π ± SD ($\times 10^{-3}$)	3.1 ± 0.0	2.4 ± 0.0	4.7 ± 0.0	2.9 ± 0.0	2.7 ± 0.0
ScIce2					
No. of genotypes	28	42	15	38	63
No. of polymorphisms	19	29	16	28	7
No. of haplotypes (private)	12 (8)	13 (9)	4 (0)	13 (7)	11 (6)
Hd± SD	0.84 ± 0.05	0.80 ± 0.05	0.54 ± 0.13	0.82 ± 0.05	0.74 ± 0.03
π ± SD ($\times 10^{-3}$)	12.5 ± 1.3	8.7 ± 1.4	8.0 ± 1.0	12.0 ± 1.1	8.4 ± 0.5
ScVrn1					
No. of genotypes	29	44	14	40	68
No. of polymorphisms	1	1	1	1	1
No. of haplotypes (private)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)
Hd± SD	0.11 ± 0.03	0.17± 0.07	0.26 ± 0.14	0.10 ± 0.06	0.09 ± 0.05
π ± SD ($\times 10^{-3}$)	0.4 ± 0.1	0.6 ± 0.2	0.9 ± 0.4	0.3 ± 0.2	0.3 ± 0.2

SD: standard deviation

Additional table 2 Percentage of SSR marker pairs in LD both genome-wide and chromosome-wise

	Chromosome (No. of markers)	**r^2>0.1 [%]**	**r^2>0.01 [%]**	***P*<0.01 [%]**
Genome-wide	Total (37)	0.4	26.5	20.8
Intra- chromosomal	1R (5)	0	41.7	41.7
	2R (5)	6.7	53.3	33.3
	3R (5)	0	13.3	6.7
	4R (4)	0	10.0	0
	5R (7)	0	10.7	57.1
	6R (4)	0	0	10.0
	7R (7)	0	7.1	14.3

r^2: strength of LD

Additional table 3 Allelic effect (β_{SNP}), SNP effect (% genetic variation explained), and *P*-value of 170 SNPs main effects and SNP x environment interaction effects in three phenotyping platforms

	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	%variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.
*ScCbf2*_SNP1	-0.26	27.88	**6.27E-05**[a]	**0.0149**	-0.18	1.04	0.7803	**0.0002**	0.75	0.00	0.5021	0.1888
*ScCbf2*_SNP2	0.02	0.13	0.7591	0.2142	0.18	2.23	0.7046	0.2324	-0.20	0.00	0.7925	0.9957
*ScCbf2*_SNP3	0.17	0.00	**0.0236**	**0.0399**	2.17	5.32	**0.0022**	0.4004	1.59	0.74	0.1904	0.9822
*ScCbf6*_SNP1	0.03	1.15	0.6609	**0.0086**	0.88	11.11	0.1040	0.2309	1.33	3.89	0.1492	**8.47E-05**
*ScCbf6*_SNP2	-0.04	0.00	0.4681	**0.0033**	0.51	7.03	0.3075	0.3924	0.76	2.79	0.3714	**5.93E-05**
*ScCbf6*_SNP3	-0.05	0.00	0.3833	**0.0027**	0.26	4.43	0.6162	0.2501	0.78	2.60	0.3685	**0.0114**
*ScCbf9b*_SNP1	0.05	1.85	0.3110	0.4967	-0.32	3.06	0.4807	0.6273	0.29	0.39	0.7038	0.6486
*ScCbf9b*_SNP2	0.05	1.67	0.3516	0.2186	-0.23	2.30	0.6116	0.5168	-0.05	0.09	0.9449	0.8646
*ScCbf9b*_SNP3	0.04	1.30	0.3945	0.2130	-0.22	2.11	0.6294	0.5787	-0.22	0.00	0.7756	0.8789
*ScCbf9b*_SNP4	0.04	0.00	0.4512	0.9727	-0.14	2.15	0.7584	0.3837	-0.22	0.07	0.7712	0.4284
*ScCbf9b*_SNP5	-0.04	0.00	0.5591	0.7890	0.79	5.76	0.1994	**0.0001**	-0.13	0.03	0.8995	0.1817
*ScCbf9b*_SNP6	-0.12	0.00	0.0995	0.7922	-0.28	0.67	0.6726	0.0905	-1.27	4.61	0.2407	**0.0185**
*ScCbf9b*_SNP7	-0.17	0.00	0.0726	0.1207	0.59	1.90	0.4474	**0.0016**	-1.17	1.88	0.3569	0.1894
*ScCbf9b*_SNP8	-0.43	22.81	**1.74E-05**	**0.0005**	-1.37	8.85	0.1303	0.1082	-1.16	7.98	0.4407	**1.87E-05**
*ScCbf9b*_SNP9	-0.25	22.69	**0.0005**	0.5768	-0.84	10.68	0.2000	0.0804	-0.33	0.56	0.7688	0.0553
*ScCbf9b*_SNP10	0.04	0.02	0.5407	0.0815	0.34	0.00	0.5319	0.6344	0.80	0.83	0.3883	0.3098

Gene_SNP	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.
*ScCbf9b*_SNP11	-0.28	23.77	**0.0003**	0.5515	-0.99	11.70	0.1748	0.0609	-0.80	4.29	0.5159	**0.0332**
*ScCbf9b*_SNP12	-0.23	22.35	**0.0013**	0.6290	-0.80	10.52	0.2253	0.0822	-0.72	1.42	0.5240	**0.0215**
*ScCbf9b*_SNP13	-0.24	19.09	**0.0013**	0.8087	-0.87	10.69	0.2027	**0.0400**	-0.59	1.42	0.6132	**0.0398**
*ScCbf9b*_SNP14	0.05	0.00	0.3195	0.1492	0.59	0.00	0.2086	0.3535	-0.42	0.00	0.6048	**0.0287**
*ScCbf9b*_SNP15	0.12	0.00	**0.0485**	**0.0024**	0.39	0.00	0.4522	0.3597	-0.36	0.00	0.6768	0.9599
*ScCbf9b*_SNP16	-0.28	26.24	**0.0002**	0.5530	-1.27	14.65	0.0746	0.2225	-1.14	5.05	0.3469	**0.0152**
*ScCbf9b*_SNP17	-0.43	23.52	**6.33E-06**	**0.0031**	-0.94	9.18	0.2879	0.0551	0.07	3.65	0.9652	**0.0001**
*ScCbf9b*_SNP18	-0.01	0.21	0.9008	**0.0109**	-0.07	0.39	0.8828	0.2033	-0.59	0.15	0.4528	0.2257
*ScCbf9b*_SNP19	-0.10	0.00	0.2814	0.0550	0.54	1.50	0.4763	0.2067	-1.92	10.18	0.1353	**0.0311**
*ScCbf9b*_SNP20	0.04	0.48	0.5532	**0.0014**	0.65	0.00	0.2327	0.5985	0.51	0.00	0.5727	0.9838
*ScCbf9b*_SNP21	0.03	0.92	0.6007	0.1242	0.23	0.87	0.6857	**0.0395**	-0.54	2.75	0.5794	0.5763
*ScCbf9b*_SNP22	0.03	0.41	0.6751	0.1328	0.22	0.45	0.6966	**0.0485**	-0.88	5.32	0.3510	0.2431
*ScCbf9b*_SNP23	-0.28	26.24	**0.0002**	0.5529	-1.27	14.65	0.0746	0.2225	-1.14	5.05	0.3469	**0.0152**
*ScCbf9b*_SNP24	-0.03	0.00	0.7040	**0.0426**	0.72	4.03	0.2732	**0.0092**	-1.00	3.59	0.3433	0.3635
*ScCbf9b*_SNP25	-0.26	18.52	**0.0003**	**2.16E-07**	-0.84	9.46	0.1695	0.0908	-1.11	4.51	0.2713	**1.82E-05**
*ScCbf9b*_SNP26	-0.24	8.26	**0.0004**	**0.0007**	-0.17	0.53	0.7819	0.2048	-1.94	19.93	0.0653	**6.12E-05**
*ScCbf9b*_SNP27	-0.18	16.37	**0.0002**	0.3205	0.67	0.00	0.1487	0.2952	-0.94	6.14	0.2160	**0.0222**

Gene_SNP	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.
*ScCbf9b*_SNP28	-0.16	0.00	0.0995	**0.0018**	0.81	5.62	0.3534	**0.0326**	0.53	1.68	0.7004	0.6170
*ScCbf9b*_SNP29	0.02	0.41	0.7932	**0.0410**	0.43	1.29	0.5030	0.0588	-0.98	4.00	0.3708	0.3316
*ScCbf9b*_SNP30	-0.05	0.00	0.5568	**0.0468**	0.84	0.00	0.2205	0.3659	-1.16	3.90	0.3102	0.2115
*ScCbf9b*_SNP31	-0.22	13.79	**0.0002**	0.8886	-0.14	0.81	0.7864	0.2824	2.02	3.01	**0.0222**	0.8535
*ScCbf11*_SNP1	0.18	0.00	0.0523	0.9963	1.82	0.00	**0.0456**	**0.0224**	-0.64	2.01	0.6609	0.4213
*ScCbf11*_SNP2	0.18	0.00	0.0523	0.9963	1.82	0.00	**0.0456**	**0.0224**	-0.64	2.01	0.6609	0.4213
*ScCbf11*_SNP3	0.18	0.00	0.0523	0.9963	1.82	0.00	**0.0456**	**0.0224**	-0.64	2.01	0.6609	0.4213
*ScCbf11*_SNP4	0.18	0.00	0.0523	0.9963	1.82	0.00	**0.0456**	**0.0224**	-0.64	2.01	0.6609	0.4213
*ScCbf11*_SNP5	0.18	0.00	0.0523	0.9963	1.82	0.00	**0.0456**	**0.0224**	-0.64	2.01	0.6609	0.4213
*ScCbf11*_SNP6	0.18	0.00	0.0523	0.9963	1.82	0.00	**0.0456**	**0.0224**	-0.64	2.01	0.6609	0.4213
*ScCbf11*_SNP7	0.18	0.00	0.0523	0.9963	1.82	0.00	**0.0456**	**0.0224**	-0.64	2.01	0.6609	0.4213
*ScCbf11*_SNP8	0.06	0.00	0.5495	0.7244	1.57	0.00	0.0760	**0.0029**	-1.04	3.93	0.4689	0.2861
*ScCbf11*_SNP9	0.10	0.00	0.2802	0.9748	1.30	0.00	0.1314	**0.0022**	-1.09	4.73	0.4359	0.3691
*ScCbf11*_SNP10	0.06	0.00	0.5495	0.7244	1.57	0.00	0.0760	**0.0029**	-1.04	3.93	0.4689	0.2861
*ScCbf11*_SNP11	0.06	0.00	0.5495	0.7244	1.57	0.00	0.0760	**0.0029**	-1.04	3.93	0.4689	0.2861
*ScCbf11*_SNP12	0.10	0.00	0.2802	0.9748	1.30	0.00	0.1314	**0.0022**	-1.09	4.73	0.4359	0.3691
*ScCbf11*_SNP13	-0.08	4.06	0.1882	0.0365	-0.52	7.64	0.3597	**0.0143**	1.11	5.85	0.2496	0.8322

	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	0.4689	SNP x envir.
*ScCbf11*_SNP14	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP15	0.06	0.00	0.5495	0.7243	1.57	0.00	0.0760	**0.0029**	-1.04	3.93	0.4689	0.2861
*ScCbf11*_SNP16	0.06	0.00	0.5495	0.7243	1.57	0.00	0.0760	**0.0029**	-1.04	3.93	0.4689	0.2861
*ScCbf11*_SNP17	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP18	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP19	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP20	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP21	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP22	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP23	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP24	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP25	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP26	-0.06	3.46	0.3113	0.0883	-0.29	5.80	0.6062	**0.0302**	1.13	7.64	0.2403	0.7648
*ScCbf11*_SNP27	-0.04	1.68	0.5515	**0.0364**	0.06	0.40	0.9196	**0.0394**	1.58	8.13	0.1013	0.9757
*ScCbf12*_SNP1	-0.32	0.00	**2.01E-05**	**0.0026**	-1.79	0.00	**0.0142**	0.6855	-0.64	0.00	0.6029	0.0940
*ScCbf12*_SNP2	-0.33	1.33	**3.49E-06**	**0.0008**	-2.01	0.00	**0.0037**	0.5858	-0.28	0.00	0.8069	0.0520
*ScCbf12*_SNP3	-0.36	0.00	**1.01E-05**	**5.52E-08**	-1.36	1.40	0.0861	0.5685	0.23	2.81	0.8637	**0.0023**

	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	0.4501	SNP x envir.
*ScCbf12*_SNP4	-0.30	4.36	**1.01E-05**	**0.0001**	-1.69	0.00	**0.0138**	0.3444	-0.60	0.00	0.5964	**0.0266**
*ScCbf12*_SNP5	0.17	12.53	**0.0059**	**0.0348**	-0.11	1.02	0.8503	**0.0491**	-0.36	5.91	0.7158	0.1559
*ScCbf12*_SNP6	-0.31	0.00	**0.0030**	**4.02E-05**	-1.44	0.00	0.1278	0.3203	3.89	28.84	**0.0140**	**1.48E-06**
*ScCbf12*_SNP10	-0.14	0.00	0.0882	0.0581	-0.58	2.48	0.4686	0.8583	-0.36	0.00	0.1971	0.7435
*ScCbf12*_SNP11	-0.20	0.00	**0.0137**	0.1291	-1.14	3.02	0.1498	0.8404	-1.69	0.00	0.1584	0.7643
*ScCbf12*_SNP12	-0.23	2.10	**0.0046**	0.1511	-1.30	4.07	0.0989	0.5391	-1.84	0.00	0.1093	0.9000
*ScCbf12*_SNP13	-0.21	1.88	**0.0110**	0.1703	-1.45	2.84	0.0723	0.7805	-2.12	0.00	0.3438	0.6859
*ScCbf12*_SNP14	-0.19	0.64	**0.0323**	**0.0265**	-1.29	7.40	0.1366	0.5959	-1.33	0.00	0.6387	0.7492
*ScCbf12*_SNP15	-0.23	6.92	**0.0017**	**0.0464**	-0.65	0.66	0.3550	0.7029	-0.56	0.00	0.4852	0.6602
*ScCbf12*_SNP16	-0.36	11.93	**6.02E-07**	**4.41E-06**	-1.97	0.00	**0.0046**	0.8984	-0.80	0.00	**0.0232**	0.3909
*ScCbf12*_SNP17	-0.26	5.91	**0.0108**	0.5331	-2.15	14.34	**0.0292**	0.0891	-3.66	6.71	0.3507	0.3935
*ScCbf12*_SNP18	-0.23	13.46	**0.0023**	**0.0037**	-1.65	7.18	**0.0274**	0.7553	-1.18	0.67	0.8362	0.6288
*ScCbf12*_SNP19	-0.28	8.15	**5.31E-05**	**5.16E-06**	-1.67	1.59	**0.0137**	0.7656	-0.23	0.00	0.9826	0.3211
*ScCbf12*_SNP20	-0.17	4.38	**0.0084**	**0.0022**	0.15	0.00	0.8056	0.4194	0.02	0.16	0.7550	0.3837
*ScCbf12*_SNP21	-0.18	0.86	**0.0321**	**0.0111**	-0.27	0.00	0.7461	0.6823	-0.43	0.00	0.5026	0.5033
*ScCbf12*_SNP22	-0.19	0.00	**0.0245**	0.4847	-1.17	5.24	0.1678	0.9416	-0.93	0.00	0.4563	0.6239
*ScCbf12*_SNP23	0.07	0.00	0.4320	**3.50E-05**	-1.41	2.70	0.0891	0.7057	1.01	2.21	0.6869	0.8709

	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	0.3696	SNP x envir.
*ScCbf12*_SNP24	-0.01	0.24	0.9506	0.4201	-2.44	17.42	**0.0056**	0.5346	-0.59	0.00	0.6869	0.4985
*ScCbf12*_SNP25	-0.13	1.09	0.0714	**0.0009**	-0.47	0.00	0.4932	0.7494	-0.94	0.21	0.3696	0.8894
*ScCbf12*_SNP26	0.11	2.29	0.0671	0.6055	-0.10	0.15	0.8559	0.9355	-1.99	11.81	**0.0369**	**0.0220**
*ScCbf14*_SNP1	-0.06	0.28	0.5108	**0.0027**	-0.56	1.65	0.4813	0.5272	-1.65	9.51	0.2086	**0.0031**
*ScCbf14*_SNP2	-0.06	0.28	0.5108	**0.0027**	-0.56	1.65	0.4813	0.5272	-1.65	9.51	0.2086	**0.0031**
*ScCbf14*_SNP3	-0.09	2.11	0.2836	**0.0096**	-0.51	1.67	0.5414	0.2355	-1.78	10.53	0.1937	**0.0032**
*ScCbf14*_SNP4	-0.06	0.28	0.5108	**0.0027**	-0.56	1.65	0.4813	0.5272	-1.65	9.51	0.2086	**0.0031**
*ScCbf14*_SNP5	-0.08	1.16	0.3520	**0.0086**	-0.37	1.53	0.6633	0.1925	-1.49	9.26	0.2846	**0.0025**
*ScCbf15*_SNP1	-0.13	5.07	**0.0391**	**2.20E-16**	-1.34	4.77	**0.0192**	0.0525	-3.15	7.26	**0.0008**	**0.0045**
*ScCbf15*_SNP2	-0.14	0.00	**0.0166**	**2.01E-12**	-1.34	1.42	**0.0112**	0.0573	-3.20	9.60	**0.0002**	**0.0002**
*ScCbf15*_SNP3	-0.22	5.71	**0.0009**	**2.20E-16**	-0.74	3.64	0.2066	0.0464	-2.32	3.91	**0.0181**	**0.0362**
*ScCbf15*_SNP4	-0.06	0.00	0.3770	**2.16E-07**	-0.93	5.99	0.1319	0.7204	-1.27	0.00	0.2226	0.2168
*ScDhn1*_SNP1	-0.14	0.24	0.1269	0.0964	0.31	0.00	0.6911	**0.0268**	-3.18	0.79	**0.0292**	0.4765
*ScDhn1*_SNP2	-0.22	0.00	**0.0143**	**0.0177**	-0.24	0.00	0.7710	**0.0208**	-3.60	1.66	**0.0198**	**0.0052**
*ScDhn1*_SNP3	-0.22	0.00	**0.0200**	**0.0006**	0.68	3.95	0.4302	0.0604	-3.36	0.00	**0.0399**	**0.0014**
*ScDhn1*_SNP4	-0.11	0.00	0.2968	0.1255	0.49	25.61	0.5610	**0.0158**	-2.49	0.00	0.1401	0.0924
*ScDhn1*_SNP5	-0.06	0.00	0.4529	0.3462	0.18	4.21	0.8051	0.0659	-1.74	0.90	0.1830	**0.0219**

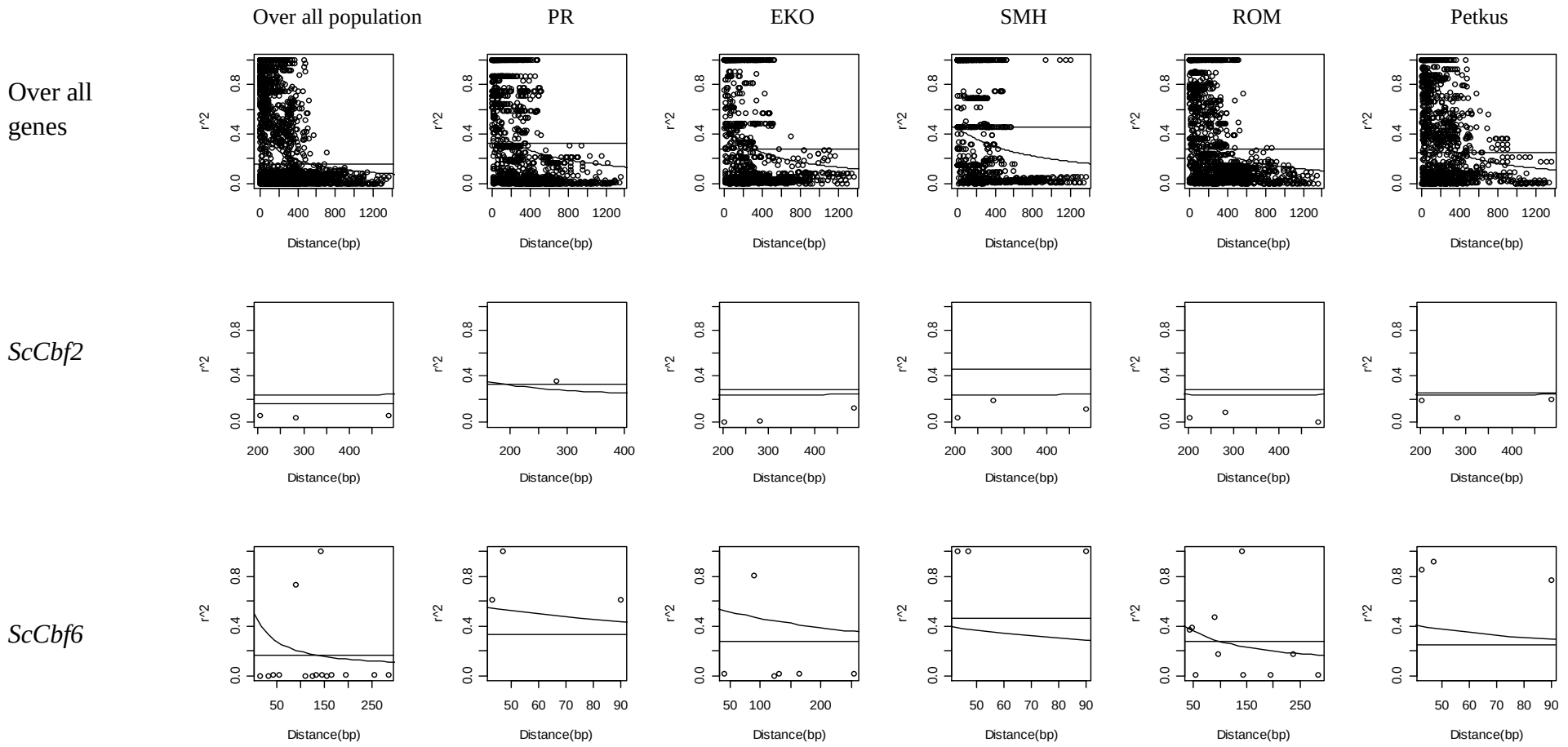
	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	0.6225	SNP x envir.
*ScDhn1*_SNP6	-0.15	0.00	0.0556	0.4082	0.43	10.19	0.5497	**0.0161**	-1.94	1.70	0.1363	**0.0455**
*ScDhn3*_SNP1	-0.03	1.02	0.7531	**0.0001**	-1.06	0.00	0.2100	**0.0465**	-1.02	7.03	0.4763	**4.15E-05**
*ScDhn3*_SNP2	0.10	0.13	0.4100	**0.0178**	-1.31	6.59	0.2101	0.0820	0.82	0.00	0.6225	0.6485
*ScDhn3*_SNP3	0.25	1.64	**0.0237**	**0.0264**	-1.50	0.00	0.1699	0.6962	0.87	0.00	0.6020	0.4408
*ScDhn3*_SNP4	0.14	2.41	0.2748	0.2470	-1.83	0.00	0.1397	0.4003	1.37	0.00	0.4661	0.5037
*ScDhn3*_SNP5	-0.04	0.00	0.4219	0.2347	0.07	1.02	0.8888	0.8242	-0.70	0.00	0.4302	0.7414
*ScDhn3*_SNP6	0.00	0.00	0.9543	**0.0395**	0.21	1.77	0.6854	0.8359	0.14	0.28	0.8737	0.5459
*ScDhn3*_SNP7	-0.03	0.00	0.5362	**0.0463**	0.12	1.01	0.8160	0.5321	-0.01	0.00	0.9871	0.3183
*ScDhn3*_SNP8	-0.02	0.00	0.6552	**0.0065**	0.28	0.81	0.5980	0.4018	0.01	0.11	0.9880	0.3412
*ScDhn3*_SNP9	0.01	0.89	0.8497	0.3555	0.00	0.14	0.9970	0.8413	0.52	2.20	0.5616	0.2293
*ScDhn3*_SNP10	0.01	0.00	0.8827	**0.0021**	-0.89	1.81	0.3017	0.9513	-1.38	4.67	0.3199	0.1326
*ScDhn3*_SNP11	0.08	0.00	0.3551	**0.0026**	-0.63	1.86	0.4641	0.9126	-0.83	2.23	0.5544	0.2959
*ScDhn3*_SNP12	-0.33	3.22	0.0699	**0.0021**	-1.20	4.34	0.3861	**0.0314**	-0.91	0.00	0.7074	0.7676
*ScDhn3*_SNP13	-0.22	3.99	0.2188	**0.0018**	-1.95	1.11	0.1810	**0.0217**	-0.11	0.03	0.9645	0.8978
*ScDhn3*_SNP14	0.09	0.00	0.3031	**0.0171**	0.30	0.00	0.7135	0.6287	0.44	0.00	0.7463	0.5049
*ScDreb2*_SNP1	-0.02	0.88	0.6928	0.3216	0.31	0.00	0.5587	0.1515	1.02	1.65	0.2375	0.5235
*ScDreb2*_SNP2	-0.08	3.53	0.3600	**0.0038**	-1.13	0.00	0.1688	0.9138	-1.17	0.46	0.4328	**0.0003**

	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	0.1648	SNP x envir.
*ScDreb2*_SNP3	-0.16	1.26	**0.0179**	0.1111	-0.69	0.00	0.2620	0.8879	-2.02	2.38	0.0611	**3.29E-05**
*ScDreb2*_SNP4	-0.05	0.00	0.4379	0.1936	-0.25	0.96	0.6359	0.4786	-1.07	1.27	0.2328	0.4509
*ScDreb2*_SNP5	-0.12	0.00	0.0629	0.3869	-0.36	0.00	0.5449	0.8406	-1.91	2.37	0.0580	**0.0009**
*ScDreb2*_SNP9	0.00	0.05	0.9575	0.5265	0.18	0.80	0.7372	0.2244	0.44	0.00	0.9530	0.7989
*ScDreb2*_SNP10	0.16	0.00	0.0801	0.1342	0.15	0.00	0.8639	0.0783	-0.09	0.74	0.7330	0.5048
*ScDreb2*_SNP11	-0.28	14.59	**0.0004**	0.3932	-0.21	0.00	0.7771	0.9675	-0.44	0.57	**0.0327**	0.9646
*ScDreb2*_SNP12	-0.13	3.13	0.1026	0.9134	0.69	3.57	0.3525	0.5809	-2.60	5.17	0.5105	**0.0006**
*ScDreb2*_SNP13	-0.05	1.66	0.6374	**0.0239**	-0.28	0.24	0.7880	0.2130	-1.13	1.80	**0.0022**	**0.0045**
*ScIce2*_SNP1	0.29	13.67	**0.0003**	**1.07E-09**	0.41	0.00	0.5777	**0.0160**	3.75	15.46	**0.0055**	**2.41E-12**
*ScIce2*_SNP2	0.32	14.71	**6.69E-05**	**6.69E-09**	0.18	0.00	0.8014	**0.0394**	3.37	10.01	**0.0055**	**4.17E-11**
*ScIce2*_SNP3	0.32	14.71	**6.69E-05**	**6.69E-09**	0.18	0.00	0.8014	**0.0394**	3.37	10.01	**0.0055**	**4.17E-11**
*ScIce2*_SNP4	0.25	12.76	**0.0019**	**2.89E-10**	0.14	0.00	0.8513	**0.0179**	3.37	10.01	**0.0093**	**4.17E-11**
*ScIce2*_SNP5	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0018**	**1.19E-08**
*ScIce2*_SNP6	0.28	5.81	**0.0017**	**7.68E-10**	1.56	0.00	0.0512	**0.0004**	4.13	12.20	**0.0014**	**9.47E-10**
*ScIce2*_SNP7	0.32	5.85	**0.0059**	**3.82E-10**	1.22	0.00	0.1285	**0.0006**	4.30	14.30	**0.0093**	**3.07E-10**
*ScIce2*_SNP8	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0093**	**1.19E-08**
*ScIce2*_SNP9	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0093**	**1.19E-08**

	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	**0.0093**	SNP x envir.
*ScIce2*_SNP10	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0093**	**1.19E-08**
*ScIce2*_SNP11	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0093**	**1.19E-08**
*ScIce2*_SNP12	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0093**	**1.19E-08**
*ScIce2*_SNP13	0.25	12.46	**0.0024**	**2.55E-11**	0.27	0.00	0.7152	**0.0068**	3.19	10.42	**0.0095**	**2.93E-11**
*ScIce2*_SNP14	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0093**	**1.18E-08**
*ScIce2*_SNP15	0.25	12.46	**0.0024**	**2.55E-11**	0.27	0.00	0.7152	**0.0068**	3.19	10.42	**0.0095**	**2.93E-11**
*ScIce2*_SNP16	0.25	12.46	**0.0024**	**2.55E-11**	0.27	0.00	0.7152	**0.0068**	3.19	10.42	**0.0095**	**2.93E-11**
*ScIce2*_SNP17	0.25	12.46	**0.0024**	**2.55E-11**	0.27	0.00	0.7152	**0.0068**	3.19	10.42	**0.0095**	**2.93E-11**
*ScIce2*_SNP18	0.30	6.05	**0.0008**	**4.88E-10**	1.43	0.00	0.0722	**0.0005**	3.44	7.30	**0.0093**	**1.19E-08**
*ScIce2*_SNP19	-0.14	0.00	**0.0074**	**1.12E-09**	0.44	3.21	0.3770	0.8994	-0.82	0.00	0.3229	0.7267
*ScIce2*_SNP20	-0.14	0.65	0.1187	0.1672	0.64	1.18	0.4897	0.2025	-0.60	0.00	0.7030	**0.0318**
*ScIce2*_SNP21	-0.11	0.00	**0.0322**	**1.23E-07**	0.44	1.50	0.3771	0.8917	-1.03	0.00	0.2070	0.6296
*ScIce2*_SNP22	-0.14	0.00	**0.0039**	**1.06E-09**	0.37	3.31	0.4433	0.7836	-1.01	0.00	0.2091	0.6386
*ScIce2*_SNP23	-0.07	0.00	0.1674	**3.28E-12**	0.54	4.49	0.2659	0.9599	-0.90	0.00	0.2627	0.6366
*ScIce2*_SNP24	0.10	0.00	0.0548	**4.69E-05**	0.01	0.23	0.9849	0.5378	0.18	0.00	0.8297	0.9665
*ScIce2*_SNP25	-0.11	0.00	**0.0433**	**2.87E-08**	0.15	0.58	0.7633	0.6315	-0.55	0.00	0.5058	0.8997
*ScIce2*_SNP26	-0.13	0.00	**0.0292**	**1.40E-06**	-0.02	0.01	0.9686	0.2441	-0.62	0.00	0.4936	0.7636

	Controlled (recovery score 0-5)				Semi-controlled (% plants with undamaged leaves)				Field (% survival)			
Gene_SNP	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	*P*- value	SNP x envir.	β_{SNP}	% variation	0.3115	SNP x envir.
*ScIce2*_SNP27	-0.12	0.00	**0.0177**	**4.38E-08**	0.21	2.40	0.6755	0.7925	-1.19	0.00	0.1508	0.7225
*ScIce2*_SNP28	-0.10	0.00	0.0600	**4.80E-11**	0.30	3.31	0.5643	0.9638	-1.00	0.00	0.2366	0.5358
*ScIce2*_SNP29	-0.11	0.00	**0.0425**	**8.65E-11**	0.12	1.12	0.8222	0.7137	-0.86	0.00	0.3115	0.8055
*ScIce2*_SNP30	-0.07	0.00	0.1582	**2.37E-09**	0.54	3.62	0.2710	0.9508	-0.71	0.00	0.3846	0.7304
*ScIce2*_SNP31	-0.11	0.00	**0.0205**	**2.42E-10**	0.52	4.91	0.2824	0.9001	-0.86	0.00	0.2792	0.7266
*ScIce2*_SNP32	-0.05	0.00	0.3082	**1.64E-08**	1.12	5.08	**0.0167**	0.7240	-0.52	0.00	0.5053	0.3258
*ScIce2*_SNP33	-0.02	0.05	0.7024	**2.25E-10**	0.80	11.03	0.0711	**0.0345**	-0.99	0.00	0.1708	0.2416
*ScIce2*_SNP34	0.00	0.00	0.9524	**2.46E-09**	1.02	12.75	**0.0218**	**0.0336**	-0.84	0.00	0.2475	0.2184
*ScIce2*_SNP35	-0.01	0.01	0.8347	**5.01E-10**	0.44	4.39	0.3415	**0.0171**	-0.93	0.00	0.2166	0.6625
*ScIce2*_SNP36	-0.03	0.00	0.4900	**3.57E-07**	0.95	8.54	**0.0328**	**0.0111**	-0.76	0.00	0.2996	0.2056
*ScIce2*_SNP37	0.03	0.00	0.4396	**3.49E-08**	1.28	10.00	**0.0035**	0.0871	-0.60	0.00	0.4055	0.4698
*ScVrn1*_SNP1	0.04	0.00	0.6560	0.2059	-1.01	0.06	0.2520	0.4924	-0.64	0.16	0.6600	0.7634

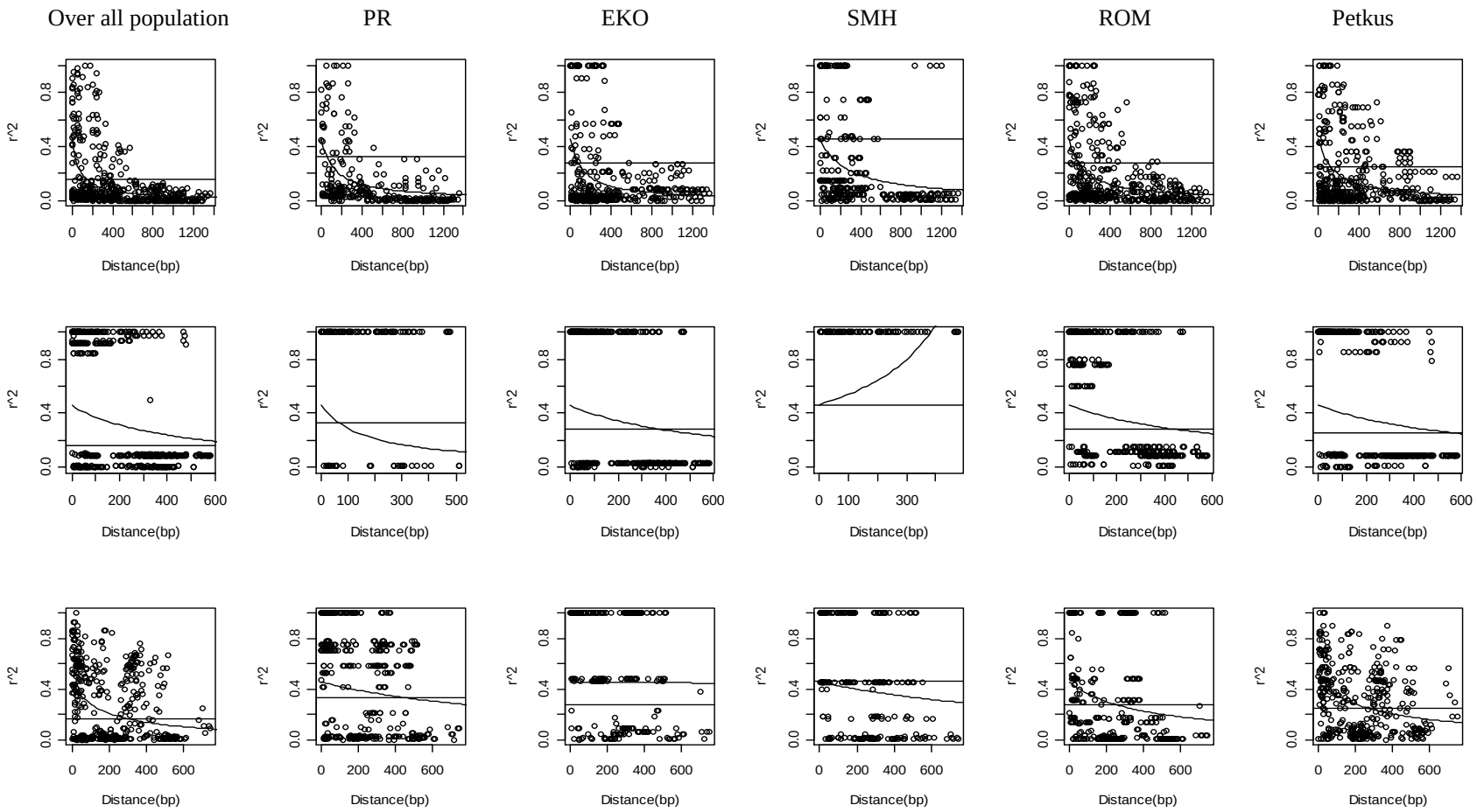
[a]*P*-values < 0.05 are printed in bold



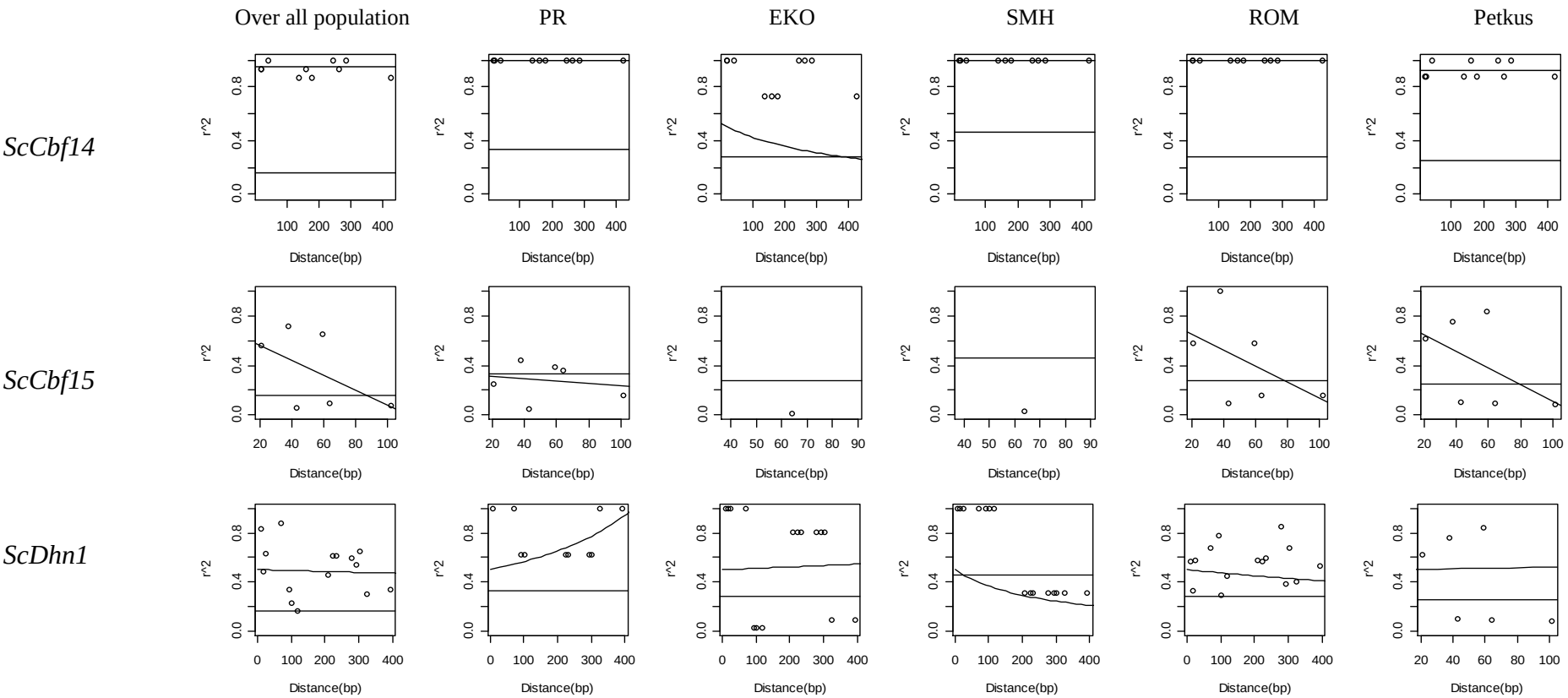
Additional figure 1 Scatterplots of pairwise distances and LD estimated by r^2 between all SNPs (MAF > 5%) in 12 candidate genes over all populations and in individual populations.

The non-linear fitting curve of the mutation-recombination-drift model is shown. Thresholds for LD are indicated by a horizontal solid line.

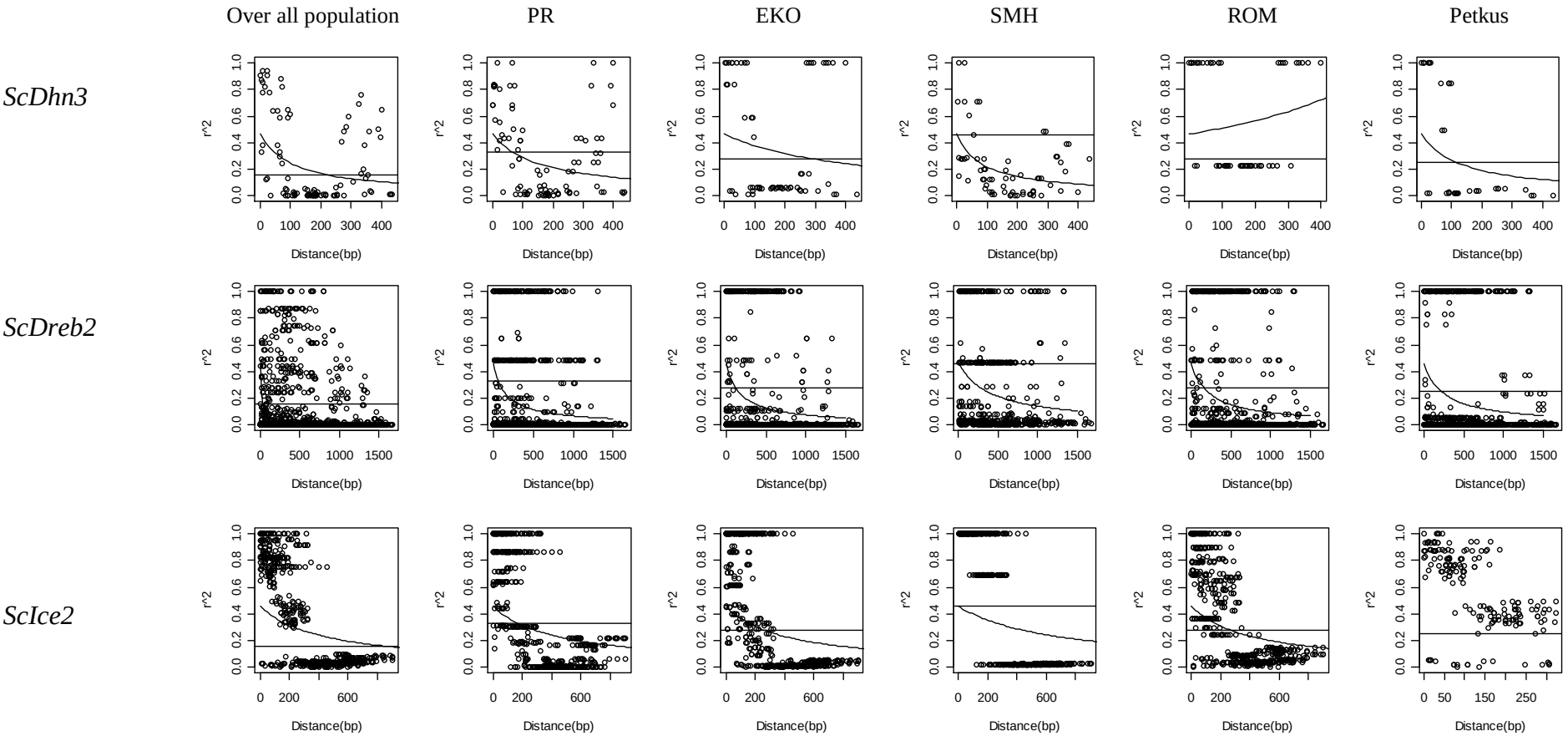
Additional figure 1 (continued)



Additional figure 1 (continued)



Additional figure 1 (continued)



9. Curriculum Vitae

PERSONAL INFORMATION

Family name	Li
First name	Yongle
Gender	Male
Birth data	07.06.1981
Address	Km. 45, Carretera México-Veracruz El Batán, Texcoco CP 56130, México
Telephone	0052(55) 5804-2004 ext.2289
E-mail	Yo.li@cgiar.org

EDUCATION

• 01/2008-10/2011 PhD candidate in plant breeding, Technische Universität München

Thesis: Association analysis of frost tolerance in rye (*Secale cereale* L.) using candidate gene polymorphisms and phenotypic data from controlled, semi-controlled, and field phenotyping platforms

Supervisor: Prof. Dr. Chris-Carolin Schön

• 10/2005-11/2007 M.Sc in environmental protection & agricultural food production,

University of Hohenheim, Germany

Thesis: Genetic diversity and linkage disequilibrium of two homologous genes to maize *D8*: Sorghum *SbD8* and pearl millet *PgD8*

Supervisor: Prof. Dr. Albrecht E. Melchinger

• 2000-2004 B.Sc in biotechnology, Guizhou University, China

Thesis: Screening heterotrophic bacteria to decompose organic pollutant for Huaxi river

Supervisor: Prof. Dr. He Xie

10. List of publications

Li Y, Boeck A, Haseneyer G, Schön CC, Ankerst D, Korzun V, Wilde P, Bauer E (2011) Association of twelve candidate genes with frost tolerance in rye under controlled, semi-controlled and field conditions. BMC Plant Biology 11:146

Li Y, Haseneyer G, Schön CC, Ankerst D, Korzun V, Wilde P, Bauer E (2011) High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (*Secale cereale* L.) genes involved in frost response. BMC Plant Biology 11:6

Li Y, Bhosale S, Haussmann BIG, Stich B, Melchinger AE and Parzies HK (2010) Genetic diversity and linkage disequilibrium of two homologous genes to maize *D8*: Sorghum *SbD8* and pearl millet *PgD8*. Journal of Plant Breeding and Crop Science 2(5) 117-128