Technische Universität München
Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Lehrstuhl für Pflanzenzüchtung

Identification of genes under differential selective pressure in temperate maize

Sandra Bettina Unterseer

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 Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:	Univ.-Prof. Dr. Brigitte Poppenberger-Sieberer
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Chris-Carolin Schön
	2. Univ.-Prof. Dr. Aurélien Tellier
	3. Univ.-Prof. Dr. Arthur Korte

Die Dissertation wurde am 08.12.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 21.03.2017 angenommen.

Content

List of Figures

List of Tables

List of Abbreviations

BSSS	Iowa Stiff Stalk Synthetic
CLR	Composite likelihood ratio (test)
F_{ST}	Fixation index
GWAS	Genome-wide association study
H_{norm}	Fay and Wu´s normalized *H*
IL	Introgression library
Indel	Insertion / Deletion
kb	Kilobase pairs
LD	Linkage disequilibrium
LSC	Lancaster Sure Crop
M	Million
Mb	Megabase pairs
Non-BSSS	Non Iowa Stiff Stalk Synthetic
OTV	Off-target variant
PHR	PolyHighResolution
π	Nucleotide diversity
QTL	Quantitative trait locus / loci
SNP	Single nucleotide polymorphism
TD	Tajima´s *D*
US	United States of America
50 k array	Illumina® MaizeSNP50 BeadChip
600 k array	Affymetrix® Axiom® Maize Genotyping Array

Publications included in this thesis

Unterseer et al. (2014)

Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H, Bertani C, Davassi A, Mayer KFX, Schön C-C (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15:823, doi: 10.1186/1471-2164-15-823.

Abstract

Background

High density genotyping data are indispensable for genomic analyses of complex traits in animal and crop species. Maize is one of the most important crop plants worldwide, however a high density SNP genotyping array for analysis of its large and highly dynamic genome has not been available so far.

Results

We developed a high density maize SNP array composed of 616,201 variants (SNPs and small indels). Initially, 57 M variants were discovered by sequencing 30 representative temperate maize lines and then stringently filtered for sequence quality scores and predicted conversion performance on the array resulting in the selection of 1.2 M polymorphic variants assayed on two screening arrays. To identify high-confidence variants, 285 DNA samples from a broad genetic diversity panel of worldwide maize lines including the samples used for sequencing, important founder lines for European maize breeding, hybrids, and proprietary samples with European, US, semi-tropical, and tropical origin were used for experimental validation. We selected 616 k variants according to their performance during validation, support of genotype calls through sequencing data, and physical distribution for further analysis and for the design of the commercially available Affymetrix® Axiom® Maize Genotyping Array. This array is composed of 609,442 SNPs and 6,759 indels. Among these are 116,224 variants in coding regions and 45,655 SNPs of the Illumina® MaizeSNP50 BeadChip for study comparison. In a subset of 45,974 variants, apart from the target SNP additional off-target variants are detected, which show only a minor bias towards intermediate allele frequencies. We performed principal coordinate and

admixture analyses to determine the ability of the array to detect and resolve population structure and investigated the extent of LD within a worldwide validation panel.

Conclusion

The high density Affymetrix® Axiom® Maize Genotyping Array is optimized for European and American temperate maize and was developed based on a diverse sample panel by applying stringent quality filter criteria to ensure its suitability for a broad range of applications.

Contribution

The candidate made major contributions to: development of the variant selection strategy; analysis of genotyping data from screening arrays and optimization of the final 600 k array; creation of all figures and tables; writing of the manuscript and revision of the paper.

Unterseer et al. (2016)

Unterseer S, Pophaly SD, Peis R, Westermeier P, Mayer M, Seidel MA, Haberer G, Mayer KFX, Ordas B, Pausch H, Tellier A, Bauer E, Schön C-C (2016) A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. Genome Biology 17:137; doi: 10.1186/s13059-016-1009-x.

Abstract

Background

Dent and Flint represent two major germplasm pools exploited in maize breeding. Several traits differentiate the two pools, like cold tolerance, early vigor and flowering time. A comparative investigation of their genomic architecture relevant for quantitative trait expression has not been reported so far. Understanding the genomic differences between germplasm pools may contribute to a better understanding of the complementarity in heterotic patterns exploited in hybrid breeding and of mechanisms involved in adaptation to different environments.

Results

We perform whole-genome screens for signatures of selection specific to temperate Dent and Flint maize by comparing high-density genotyping data of 70 American and European Dent and 66 European Flint inbred lines. We find 2.2 % and 1.4 % of the genes are under selective pressure, respectively, and identify candidate genes associated with agronomic traits known to differ between the two pools. Taking flowering time as an example for the differentiation between Dent and Flint, we investigate candidate genes involved in the flowering network by phenotypic analyses in a Dent–Flint introgression library and find that the Flint haplotypes of the candidates promote earlier flowering. Within the flowering network, the majority of Flint candidates are associated with endogenous pathways in contrast to Dent candidate genes, which are mainly involved in response to environmental factors like light and photoperiod. The diversity patterns of the candidates in a unique panel of more than 900 individuals from 38 European landraces indicate a major contribution of landraces from France, Germany, and Spain to the candidate gene diversity of the Flint elite lines.

Conclusions

In this study, we report the investigation of pool-specific differences between temperate Dent and Flint on a genome-wide scale. The identified candidate genes represent a promising source for the functional investigation of pool-specific haplotypes in different genetic backgrounds and for the evaluation of their potential for future crop improvement like the adaptation to specific environments.

Contribution

The candidate made major contributions to: conceiving the study and discussion of results; analysis of genotypic data of elite lines and landraces; analysis of sequence data; investigation of the introgression library data; creation of all figures and tables; writing of the manuscript and revision of the paper.

1 Introduction

Maize (*Zea mays* ssp. *mays* L.) is one of the most important crops worldwide and represents an intensively studied organism. The objective of this study was the identification of genomic regions under differential selective pressure in two major temperate maize germplasm pools and the investigation of candidate genes underlying phenotypic variation. The species maize as well as population genetic approaches for the detection of signatures of selection are introduced in section 1.1, followed by the outline of the thesis given in section 1.2.

1.1 Background

Maize is an important source for food, livestock feed and industrial products and can be cultivated in a wide range of environmental conditions, for example in the Americas from Canada to Chile. The success of maize can be summarized by two key factors: i) a tremendous genetic diversity that facilitated its adaptation to various climates, and ii) the establishment of divergent heterotic groups in hybrid breeding, leading to an enormous increase in yield.

The genomic diversity of maize has been shaped by adaptation and selection since its domestication. Maize was domesticated from its wild ancestor teosinte about 9,000 years ago in Mexico (Matsuoka et al. 2002; van Heerwaarden et al. 2011) by stringent selection for naturally occurring maize-like phenotypes, e.g. plants with shortened lateral branches tipped by female ears (Piperno et al. 2014). These changes in phenotypes were accompanied by considerable changes of the genetic, transcriptional and structural architecture of maize (Hufford et al. 2012; Matsuoka et al. 2002; Piperno et al. 2014; Swanson-Wagner et al. 2012). Subsequent to domestication, maize landraces were subjected to artificial selective pressure for important agronomic traits such as yield and resistance to biotic and abiotic stresses, thus giving rise to improved maize lines. Based on a survey of the genetic composition of 774 genes, which were compared between modern maize lines and teosinte, it has been estimated that 2-4% of all genes have been under artificial selective pressure in maize, thus corresponding to roughly 1,000 genes (Wright et al. 2005). In a comparative study based on teosinte accessions, landraces and improved maize lines, it was shown that landraces retained more than 80% of the genetic diversity of the wild ancestor, which is more than in other crop species (Hufford et al. 2012). The study also revealed that the effect of domestication on the genome-wide pattern of diversity in modern maize lines was stronger compared to the impact of improvement and that elite lines retained more than 98% of the genetic diversity of landraces (Hufford et al. 2012). Furthermore, a recent study in maize showed that the majority of quantitative trait loci (QTL) exhibit phenotypic effects dependent on a given environmental condition (Millet et al. 2016). Thus, the high proportion of retained genetic diversity in the elite lines could be explained by the improvement of locally adapted landraces for different agronomic traits, which also paved the way for modern hybrid breeding.

A key step in corn production was the discovery of heterotic effects in maize hybrids, which arise from the combination of inbred lines from different germplasm (Shull 1909). Two major heterotic pools exploited in European hybrid maize breeding are Dent and Flint, with their names referring to different kernel phenotypes (Smith et al. 2004). Dent lines have characteristic indented kernels with high soft starch content, whereas Flint lines have kernels with a thick, hard, and vitreous outer layer. Worldwide, many hybrid breeding programs exploit heterotic effects between different pools within Dent, like in the US. The modern US Corn Belt Dent germplasm consists of multiple heterotic pools, which can be classified into Iowa Stiff Stalk Synthetic (BSSS), Iodent, and Lancaster Sure Crop (LSC) as well as a group of lines with diverse background that is referred to as non-BSSS herein (Bennetzen and Hake 2009; Mikel and Dudley 2006). To maximize hybrid performance in breeding schemes, ongoing selective pressure resulted in increased divergence between these germplasm pools associated with an increase of genetic similarity within each pool (van Heerwaarden et al. 2012). The majority of modern US Dent germplasm traces back to a small number of founder lines including the genome reference sequence line B73 (BSSS), PH207 (Iodent), and Mo17 (LSC; Mikel and Dudley 2006). This is exemplarily depicted in Figure 1 based on pedigree information that was available for a set of Dent lines included in this study. Most of the Dent lines investigated in this thesis were US Corn Belt Dent, whereas most of the Flint lines were derived from European breeding programs. European maize has a diverse background due the introduction of maize from different parts of America. Maize was introduced to Europe at the end of the 15th century, when Columbus brought subtropical maize from the Caribbean Islands to Southern Spain, followed by travellers importing Northern Flint from Canada to Northern France (Figure 2; Dubreuil et al. 2006). Northern Flint reached very high latitudes, which required the adaptation to cooler and shorter vegetation periods (Brown and Anderson 1947). It was a major progenitor of maize in most European regions enabling its rapid adaptation to European climates (Bouchet et al. 2013; Rebourg et al. 2003). Therefore, especially in cooler regions of Central Europe, breeding programs exploit heterotic effects between Dent lines tracing back to US

Corn Belt Dent and Flint lines, with Flint contributing early vigour and good cold tolerance and Dent contributing high productivity to the hybrids (Schmidt 2003; Schnell 1992).

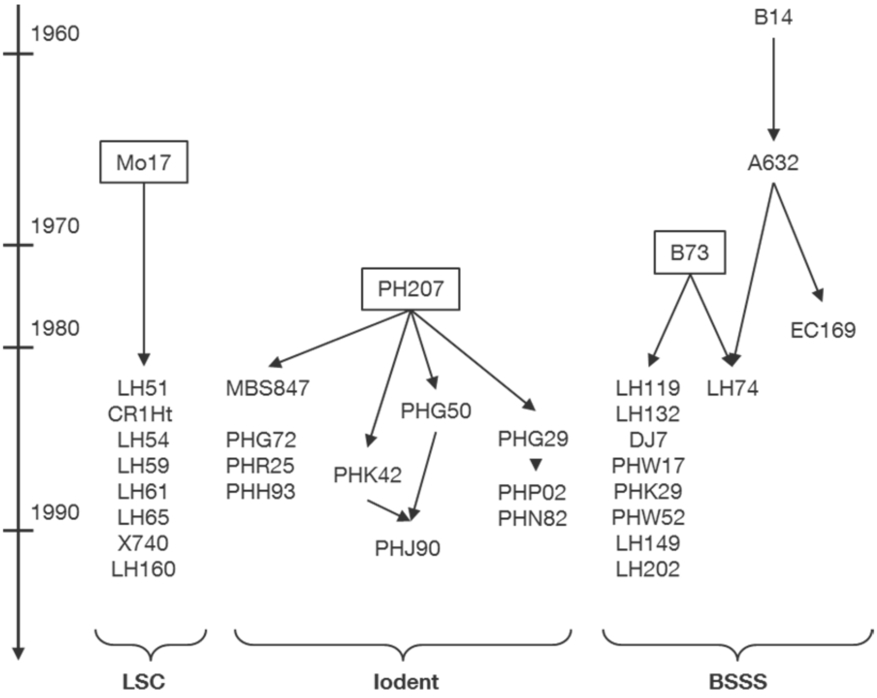

Figure 1: Contribution of important founder lines (black boxed) to the pedigree of Dent lines under study. Founder lines and their progenies are grouped along the vertical line according to the respective year of release.

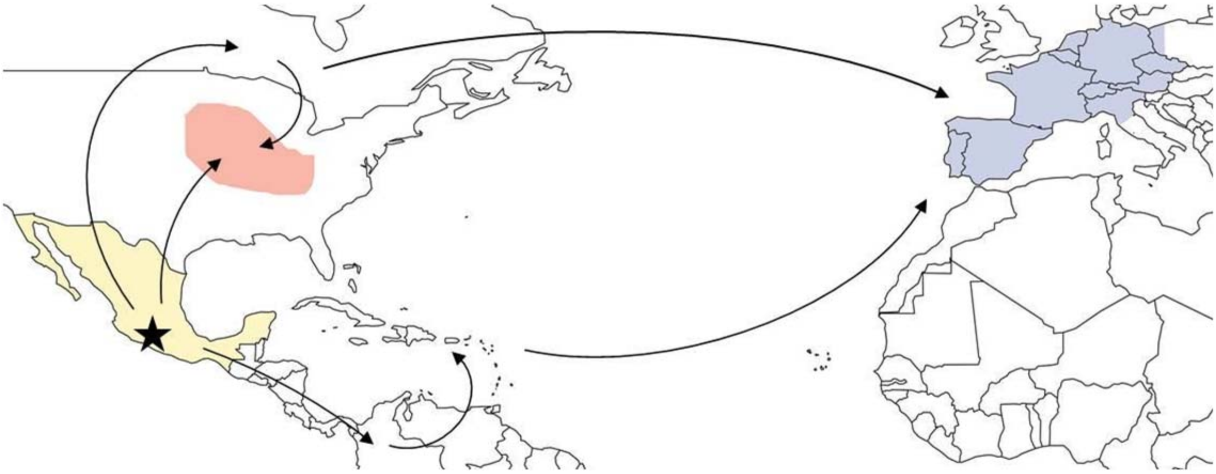

Figure 2: Historical expansion routes of maize and geographic distribution of the maize material investigated in this thesis. Arrows represent hypothetical expansion routes of maize during its historical spread along the Americas from its centre of origin in Mexico (star; modified after Tenaillon and Charcosset 2011). Yellow colour indicates the geographic distribution of the majority of lines from (sub)tropical regions. Material of most of the Dent lines was derived from the US Corn Belt region (red colour), whereas the majority of the Flint lines as well as the landraces were obtained from Europe (blue colour).

For hybrid breeding, profound knowledge of genes involved in heterotic effects would be valuable to maximize heterosis by targeted crossing of beneficial allelic combinations. More than one century has passed since the first description of heterosis by Shull (Shull 1908), and research efforts still focus on the investigation of molecular mechanisms underlying this phenomenon (Baranwal et al. 2012; Birchler et al. 2010; Feng et al. 2015; Kaeppler 2012; Schön et al. 2010). Enormous technological and bioinformatic advancements enabled the creation of comprehensive whole-genome sequencing data, which provided insights into the flexible and dynamic genome of maize. The reference sequence genome of the maize inbred line B73 was published in 2009 (Schnable et al. 2009). It revealed that almost 85% of the sequence was composed of transposable elements (Schnable et al. 2009; Wei et al. 2009), originating from an ancient explosion of repetitive DNA (Du et al. 2006). By analysing additional 27 inbred lines as part of the maize HapMap project (Gore et al. 2009), it was estimated that the B73 reference sequence assembly might contain only 70% of the existing gene space of maize. In 2012, the HapMap2 dataset was published comprising 103 lines of maize and its ancestor teosinte, reporting varying genome sizes and extensive structural variations (Chia et al. 2012). These findings advanced the concept of a pan-genome, which comprises genomic segments common to all lines and dispensable segments that can be line-specific or partially shared between lines (Morgante et al. 2007). Recent studies investigated the relevance of the pan-genome and pan-transcriptome for phenotypic trait variation (Hirsch et al. 2014b; Lu et al. 2015; Springer et al. 2009; Swanson-Wagner et al. 2010) and showed their contribution to heterotic effects exploited in hybrid breeding (Jin et al. 2016).

Knowledge-driven improvement of breeding strategies requires a comprehensive understanding of genes under selection. This includes the investigation of maize germplasm on a genomic level to elucidate how pool-specific selection shaped its genomic diversity. Selection creates specific patterns of diversity in the genome and these local signatures can be used for the detection of regions under selection (Nielsen 2005). In a hard sweep scenario, a new mutation with favourable effects on the phenotype will rise in frequency in a given population up to fixation. As the favourable allele spreads through the population, long, unbroken haplotypes flanking the selected allele are transmitted as the dispersal of the favourable allele is faster than recombination is able to break down linkage disequilibrium (LD), the non-random association of alleles at two or more loci. Therefore, not only the selected site, but also the surrounding sites are characterized by low nucleotide diversity and extreme allele frequencies, as the unfavourable allele and its adjacent sites will gradually be replaced. Based on these sweep characteristics, several methods have been proposed to detect signatures of selection (Vitti et al. 2013) and methods often are combined to minimize the possibility of false positives (Long et al. 2013; Pickrell et al. 2009; Qanbari et al. 2011). Some tests focus on the detection of local changes in allele frequencies, as a sweep is characterized by sequence variants with high derived allele frequencies (Fay and Wu 2000; Zeng et al. 2006) and a surplus of rare alleles due to the independent occurrence of new mutations in the regions flanking the selected site (Tajima 1989). Other tests are designed to detect recent sweeps based on long haplotype blocks with high LD surrounding the selected site (Sabeti et al. 2002; Voight et al. 2006). A third class of tests is based on population differentiation (Fariello et al. 2013; Lewontin and Krakauer 1973). The assumption underlying the latter class of tests is that in case of a group-specific sweep the resulting local changes in allele frequency within this group are associated with an increase in allele frequency differences between groups (Beaumont 2005).

For the detection of hard sweeps, especially frequency-dependent metrics are powerful as the allele frequency spectrum contains most information in case of this sweep scenario (Kim and Nielsen 2004). However, the identification of hard sweeps can be hampered in some circumstances. Adjacent hard sweeps can be disguised as partial or soft sweeps due to the interference of neighbouring selected alleles rising in frequency (Schrider et al. 2015). Soft sweeps refer to different sweep scenarios that share the selection of several haplotypes at varying frequencies. As a result, soft sweeps can partially retain the original variation at linked neutral sites (Hermisson and Pennings 2005). Soft sweeps can arise in case of i) selection on standing genetic variation, where a favourable allele was selected that had existed before the onset of selective pressure, ii) selection of one of several favourable alleles, and iii) parallel selection of favourable alleles in structured populations (Hermisson and Pennings 2005; Innan and Kim 2004; Messer and Petrov 2013; Przeworski et al. 2005). If a population is structured, for example due to local adaptation or varying artificial selective pressure, also hard sweeps can be masked, as the allele frequency distribution will be biased towards a higher fraction of intermediate allele frequencies. This can resemble soft sweeps and bias furthermore the estimation of the differentiation level between populations (Beaumont and Balding 2004; Muirhead 2001). Additionally, the identification of selective sweeps can be hampered by demography, such as changes in population size over time (Wright and Gaut 2005). For example in case of recent population growth, expansion is associated with an increase of segregating sites with low allele frequencies, which gives rise to patterns that might be interpreted as signals of positive selection (Maruyama and Fuerst 1984). Population size changes can be modelled to avoid demographic effects as confounding factors in population genetic analyses (Schraiber and Akey 2015; Wu et al. 2014). However, theory-driven population genetic models imply many assumptions including homogeneous, randomly mating populations, constant effective population size and constant selection intensity over time and space, as well as absence of interfering effects such as background selection and recurrent mutations at the same position in the genome. Most of these assumptions are violated in populations under artificial selection as in case of maize. Its demographic history has been modelled recently with respect to domestication (Beissinger et al. 2016), though a comprehensive demographic model has not been inferred so far that is considering its complex breeding history with multiple genetic bottlenecks and admixture events, drift and population structure. In spite of these challenges, selective sweeps have been detected in different species by comparing the allelic composition of defined genetic regions with the genomic background (Horton et al. 2012; Hufford et al. 2012; Xie et al. 2015). Candidate genes were identified and phenotypic effects of candidate genes were confirmed for example by QTL mapping or genome-wide association studies. However, a direct validation of the phenotypic effects of identified candidate genes has been rarely reported in the literature so far (Hufford 2016).

1.2 Outline

Different types of data can be used for the investigation of genomic differences between maize germplasm. Whole-genome sequencing data provide high-density information, which is favourable for population genetic analyses for example. In maize, the identification of sequence variants for genomic analyses faces specific challenges due to its evolutionary history and the high variability of its genome. As an ancient polyploid species, the maize genome is characterized by numerous duplicated chromosomal regions giving rise to paralogous sequences (Ahn and Tanksley 1993; Schnable et al. 2009; Schnable et al. 2011). Furthermore, the high amount of transposable elements, paralogs, and structural variation, including copy number and presence/absence variation, represents a challenge for variant identification due to ambiguous sequence read mapping results (Chia et al. 2012; Schnable et al. 2009; Springer et al. 2009). As sequence coverage, and thus data quality, depends on the trade-off between costs and the number of sequenced samples, the genotyping by sequencing technology is increasingly used (Elshire et al. 2011). With lower costs compared to whole-genome shotgun sequencing, this approach has been applied to more than 60 k maize samples up to now, and for the latest release of the HapMap project (Bukowski et al. 2015). The approach is based on the sequencing of genomic regions targeted by restriction enzymes (Elshire et al. 2011). However, data generated by the genotyping by sequencing technology are characterized by non-uniform distribution of sequence reads, low variant call rates and a substantial amount of missing calls (Beissinger et al. 2013; Romay et al. 2013).

On the other hand, genotyping arrays represent complexity-reducing alternatives to whole-genome sequencing efforts (Voss-Fels and Snowdon 2016). They offer a high-throughput and cost-efficient possibility to gain high-quality genomic information with low bioinformatic demands. At the beginning of this thesis, a commercial mid-density genotyping array was available for maize, which included 56,112 SNPs (Ganal et al. 2011). However, with respect to the genome size of maize and its high level of diversity, higher marker density was desirable. Moreover, it was reported that LD decays rapidly in diverse maize panels (Lu et al. 2011; Romay et al. 2013; Yan et al. 2009), thus emphasising the requirement of higher marker densities than so far available on genotyping arrays for dissecting complex agronomic traits using QTL mapping or genome-wide association studies (GWAS). To gain higher genome-wide coverage of sequence variants, a new high density genotyping array was developed as part of this thesis based on the available B73 reference sequence and whole-genome sequence data of a panel of representative European and US maize lines. Based on a stringent multi-step filtering approach and the validation of a filtered set of sequence variants by genotyping a broad genetic diversity panel of worldwide maize lines, high-confidence variants were selected for the development of a 600 k genotyping array. The array was optimized for European and American temperate maize and is well suited for fine-mapping of genomic regions, haplotype construction, detection of marker-trait associations, and first insights into the genomic composition of large diversity panels. The discovery of sequence variants, the variant filtering process, the design of the final genotyping array, and its exemplary application for resolving population structure and LD in a panel of temperate Dent and Flint inbred lines were described in Unterseer et al. (2014).

The divergence of the Dent and Flint germplasm groups has been described in diversity studies based on molecular markers (Dubreuil et al. 2006) and also in genetic studies mapping QTL underlying agronomic traits. In a recent study that utilized Dent and Flint nested association mapping populations (Bauer et al. 2013), little overlap of QTL was found for five complex traits between the two pools (Giraud et al. 2014). Thus, this thesis aimed at identifying genes under differential selective pressure in temperate maize to shed light on the genomic differentiation between temperate Dent and Flint germplasm and to unravel candidate genes for crop improvement. In maize, different approaches have been applied to investigate its genetic and phenotypic diversity, evolutionary processes shaping its genome, the genetic base of quantitative traits, and heterotic effects exploited in hybrid breeding.

Effects of selection were studied in case of individual (Vann et al. 2015; Wang et al. 2005; Wills et al. 2013; Xu et al. 2014) or a limited number of genes (Vigouroux et al. 2002; Yamasaki et al. 2005). To obtain a comprehensive view of the effects of selection on the genome, long-term divergent and recurrent selection experiments have been published (Beissinger et al. 2014; Durand et al. 2015; Hirsch et al. 2014a; Sekhon et al. 2014; Teixeira et al. 2015) and genome-wide screens were successfully applied with respect to signatures of domestication and improvement (Hufford et al. 2012; Jiao et al. 2012). However, the majority of studies focused on the US Dent pool and/or tropical maize, thus questions about genome-wide targets of differential selective pressure between Dent and Flint have not been addressed so far. Well adapted to cooler climates, the Flint pool is an integral part of European breeding programs. Moreover, the Flint germplasm pool might represent an important source of alleles associated with early vigour as well as early flowering (Brown and Anderson 1947; Rebourg et al. 2003). This could be of special relevance considering that the effects of climate change might be mitigated by shifting production areas towards higher latitudes (Lobell and Tebaldi 2014). In a comparative genomics approach, pool-specific genomic differences between Dent and Flint were investigated in Unterseer et al. (2016). Based on a combination of population genetic statistics, hundreds of candidate genes for Dent and Flint were identified. Focussing exemplarily on the flowering network in maize, its differential modulation was shown in Dent and Flint lines. Furthermore, the Flint candidate gene haplotypes were linked to phenotypic effects, thus revealing their positive impact on promoting earlier flowering time. Finally, it was demonstrated that most of the selective pressure preceded the development of modern Flint elite lines from European Flint landraces.

Here, results presented in Unterseer et al. (2014) and Unterseer et al. (2016) are complemented by additional findings relevant for the evaluation of the performed genome-wide screen for signatures of differential selective pressure in temperate Dent and Flint and the discussion of the obtained results.

2 Material and methods

As part of this thesis, a high-density maize genotyping array was developed and used for population genetic analyses based on different maize datasets (Unterseer et al. 2014; Unterseer et al. 2016). An overview of the investigated datasets is given in section 2.1, followed by the description of the genotyping array development in section 2.2. Population genetic metrics are introduced in section 2.3 and analyses of population structure and the investigation of LD in the genetic material under study are presented in section 2.4. Section 2.5 describes the identification of candidate genes under differential selective pressure in Dent and Flint and section 2.6 summarizes the investigation of these genes based on additional datasets.

2.1 Overview of datasets

Different datasets were investigated within this thesis, which are summarized in Figure 3 and will be described in the following paragraphs.

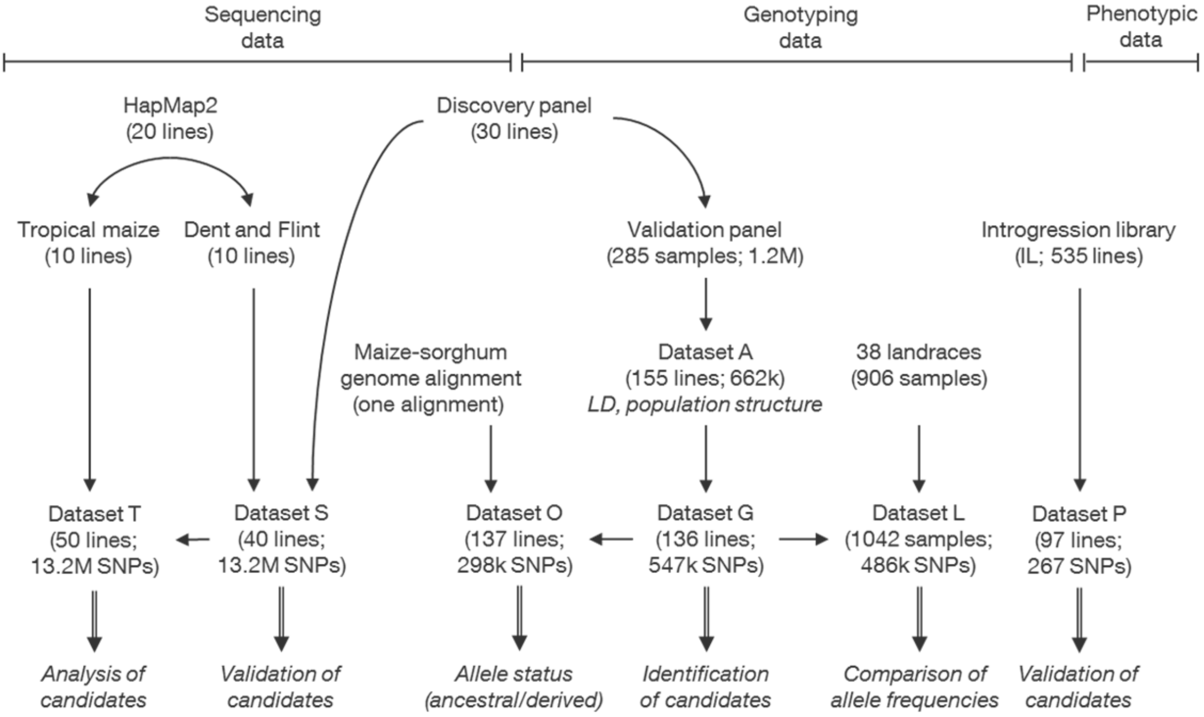
Sequencing data
Genotyping data
Phenotypic data
HapMap2 (20 lines)
Discovery panel (30 lines)
Tropical maize (10 lines)
Dent and Flint (10 lines)
Validation panel (285 samples; 1.2M)
Introgression library (IL; 535 lines)
Maize-sorghum genome alignment (one alignment)
Dataset A (155 lines; 662k) LD, population structure
38 landraces (906 samples)
Dataset T (50 lines; 13.2M SNPs)
Dataset S (40 lines; 13.2M SNPs)
Dataset O (137 lines; 298k SNPs)
Dataset G (136 lines; 547k SNPs)
Dataset L (1042 samples; 486k SNPs)
Dataset P (97 lines; 267 SNPs)
Analysis of candidates
Validation of candidates
Allele status (ancestral/derived)
Identification of candidates
Comparison of allele frequencies
Validation of candidates

Figure 3: Overview of datasets analysed in this thesis.

For the establishment of a high-density array for maize, a discovery panel was sequenced to obtain a comprehensive view of Flint- and Dent-specific genomic variation. The panel comprised 17 Flint and 13 Dent lines, which were selected to represent temperate Dent and Flint germplasm exploited in US and European breeding programs (Bauer et al. 2013). Fourteen Flint and 12 Dent lines were sequenced with on average 12-fold and three Flint and one Dent line with on average 50-fold coverage. In total, 56,938,462 variant positions were detected (Unterseer et al. 2014).

To select sequence variants for the construction of the genotyping platform, variants were filtered for quality parameters, distribution along the genome, and predicted conversion performance (Unterseer et al. 2014). A total of 1,228,505 sequence variants remained for experimental validation in the validation panel. This panel was composed of 285 DNA samples and reflected the diversity of maize with special emphasis on temperate US and European germplasm. Excluding proprietary material and replicates, genotype calls could be obtained for 129 temperate Dent and Flint inbred lines (including the 30 lines of the discovery panel), 13 tropical lines, ten doubled haploid lines from three European Flint landraces, three lines with no available pool assignment, two teosinte accessions and 23 F1 hybrids from Mendelian trios with both parental lines present in the validation panel. The best performing 616,201 variants were included on the 600 k array (Unterseer et al. 2014).

To evaluate the performance of the genotyping array concerning analyses of population structure and LD, a subset of the validation panel, dataset A, was investigated in Unterseer et al. (2014). Dataset A consisted of homozygous genotype calls of 155 public lines of the validation panel, namely 129 temperate Dent and Flint lines, 13 tropical lines, ten doubled haploid lines from three European Flint landraces, and three lines with no available germplasm assignment. For PCoA and ADMIXTURE analysis, 45,974 sequence variants classified as "off-target variants" (OTVs; for details of quality categories see section 2.2) were included with the respective genotype call of the target variant as well as the information on presence or absence of flanking variants, resulting in 616 k plus 46 k variants. For all analyses based on dataset A, indels were treated as bi-allelic SNPs and variants with ≥ 10% of missing data were excluded.

For the identification of genes under differential selective pressure between Dent and Flint as well as the investigation of population structure and haplotype blocks, dataset G was created (Unterseer et al. 2016). It comprised 136 temperate lines of dataset A and was composed of 70 Dent lines and 66 Flint lines with unambiguous germplasm assignment. Upon exclusion of monomorphic SNPs and SNPs that were designed to specifically differentiate between two Dent lines (Frascaroli et al. 2013; Ganal et al. 2011), analyses of dataset G were based on 547,412 best-quality SNPs (PHR; for details of quality categories see section 2.2).

To compare the allelic composition of candidate and non-candidate genes between elite lines and landraces, dataset L was investigated (Unterseer et al. 2016). Dataset L comprised genotype calls of dataset G and genotype calls of 906 individuals from 38 European landraces (Table 1) with 31 landraces displaying Flint-type kernels and seven landraces at least partially Dent-type kernels. Landraces were selected with the aim to reflect the genetic and phenotypic diversity of Central and Western Europe and were represented by 22 to 24 individuals each. Samples were genotyped using the 600 k array and a genotype cluster model file that was generated based on genotype calls of the validation set. Analyses were based on 486,208 SNPs, which were of best quality in elite lines and landraces.

To study derived allele frequencies, dataset O was created. It included genotype calls of dataset G and information from the maize-sorghum (*Sorghum bicolor* L.) genome alignment. The alignment was downloaded from http://pipeline.lbl.gov/downloads.shtml and was used to obtain the nucleotide in sorghum representing the ancestral maize allele. Information regarding the ancestral allele state was available for 298,388 SNPs of dataset G and used to polarize the genotype calls of dataset G with respect to sorghum (Unterseer et al. 2016).

In dataset S, candidate genes identified based on dataset G were investigated (Unterseer et al. 2016). Genotype calls of the discovery panel were combined with calls of ten temperate Dent and Flint inbred lines from the maize HapMap2 project for the respective SNPs (Chia et al. 2012; Hufford et al. 2012). Thus, dataset S was composed of genotype calls of 19 temperate Flint and 21 temperate Dent lines and included 13,246,294 bi-allelic and homozygous SNPs. SNPs were further filtered for ≤ 50.0% missing values across the 40 lines for the estimation of the level of allelic differentiation between Dent and Flint and for ≤ 50.0% missing values within germplasm pools for the calculation of gene-wise diversity.

Table 1: Landraces under study with their geographic origin (modified after Unterseer et al. 2016).

Landrace	Abbreviation	Geographic origin
Altreier	AL	Altrei, South Tyrol, Italy
Andoain	AN	Andoain, Basque Country, Spain
Barisis	BA	Barisis, Nord-Pas-de-Calais-Picardie, France
Bugard	BU	Bugard, Languedoc-Roussillon-Midi-Pyrénées, France
Castellote	CA	Castellote, Aragon, Spain
Colmar	CO	Colmar, Alsace-Champagne-Ardenne-Lorraine, France
Fleimstal	FL	Fiemme Valley, South Tyrol, Italy
Gazost	GA	Gazost, Languedoc-Roussillon-Midi-Pyrénées, France
Gelber Badischer Landmais	GB	Upper Rhine valley, Germany
Gleisdorfer	GL	Gleisdorf, Styria, Austria
Kemater Landmais	KL	Kematen, Tyrol, Austria
Knillis	KN	Styria, Austria
Krajova c29	KR	Craiova, Moravské Lieskové, Slovakia
Lacaune	LC	Lacaune, Languedoc-Roussillon-Midi-Pyrénées, France
Lalin	LL	Lalín, Galicia, Spain
Lucq de Bearn	LD	Lucq-de-Béarn, Aquitaine-Limousin-Poitou-Charentes, France
Mahndorfer	MD	Northern Germany
Maleksberger	MB	Northern Germany
Millette du Lauragais 2	ML	Lauragais, Languedoc-Roussillon-Midi-Pyrénées, France
Moncassin	MO	Moncassin, Languedoc-Roussillon-Midi-Pyrénées, France
Nostrano dell Isola	ND	Northern Italy, Italy
Oberhuber Martha	OM	Innsbruck, Tyrol, Austria
Österreichische Landsorte	OE	Upper Austria, Austria
Petkuser Ferdinand Rot	PE	Northeastern Germany
Pfarrkirchner	PF	Pfarrkirchen, Bavaria, Germany
Polnischer Landmais	PL	Poland
Rheintaler Monsheim	RM	Monsheim, Rhineland-Palatinate, Germany
Rheintaler St. Gallen	RT	St. Gallen, St. Gallen, Switzerland
Rottaler	RO	Rottal-Inn, Bavaria, Germany
Roux de Chalosse	RD	Chalosse, Aquitaine-Limousin-Poitou-Charentes, France
Santiago	SA	Santiago de Compostela, Galicia, Spain
Schindelmeiser	SC	Northeastern Germany
Sornay	SO	Sornay, Bourgogne-Franche-Comté, France
Strenzfelder	SF	Southeastern Germany, Germany
Tremesino	TR	Mediterranean Spain
Tui	TU	Tui, Galicia, Spain
Viana	VI	Viana, Galicia, Spain
Wantzenau	WA	La Wantzenau, Alsace-Champagne-Ardenne-Lorraine, France

Dataset T was investigated with respect to differences in allele frequencies of candidate and non-candidate genes between tropical and temperate maize lines. It included genotype calls of dataset S and of ten tropical lines obtained from the HapMap2 study (Chia et al. 2012), namely CML52, CML69, CML103, CML228, CML247, CML277, CML322, CML333, Ki3 and Tzi8. For analyses based on dataset T, the 13,246,294 SNPs included in dataset S were further filtered for SNPs with ≤ 50.0% missing values.

Dataset P refers to an introgression library (IL) that was investigated with respect to the effect of specific genomic segments carrying candidate genes identified based on dataset G on the phenotype (Unterseer et al. 2016). The IL was composed of 535 lines carrying genomic segments of a Flint parent in a Dent genetic background with 97 lines carrying a single Flint segment. Two field experiments were conducted in 2014 to assess flowering time, recorded as days after sowing, for the introgression lines, the two parental lines, and a check. Each experiment was laid out as an α-lattice design with two replications, except for parental lines and the check that were repeated three and five times, respectively. The genomic composition of the 97 single-segment IL lines was determined based on 267 SNPs of the Illumina® MaizeSNP50 array (50 k array; Ganal et al. 2011).

2.2 Development of the 600 k genotyping array

The Affymetrix® Axiom® Maize Genotyping Array (600 k array) was developed as a new tool to study genomic differences in maize at high density (Unterseer et al. 2014). Based on whole-genome sequence data of the 30 temperate Dent and Flint elite lines of the discovery panel, a total of 57 M variant positions was identified by mapping the generated sequence reads to the B73 reference sequence version 2 (Chia et al. 2012). A multi-step filtering approach was then applied to reduce the number of sequence variants to 1.2 M variants for experimental validation on screening arrays. This approach included filtering for sequence variants that were identified independently by two different programs, SAMtools (Li et al. 2009) and GATK (McKenna et al. 2010), and that were associated with intermediate read coverage, high mapping and SNP quality scores, as well as high predicted conversion quality scores according to the Affymetrix® Axiom® myDesign GW bioinformatics pipeline. Using a bin-based approach, the next filtering step created a set of physically equally distributed variants with a balanced representation of germplasm-specific as well as shared variants between Dent and Flint. The resulting set of 1.2 M sequence variants included 150,394 coding variants and 48,324 variants from the 50 k array (Ganal et al. 2011) and was used to genotype the validation set. To assemble a robust set of variants for designing the final 600 k array, the conversion performance of the variants was investigated based on genotype call rates, signal cluster separation, and reproducibility, polymorphism in the validation panel, and consistent Mendelian inheritance from parents to off-spring in trios. The classification of sequence variants based on signal cluster metrics will be explained below as the application of a flexible clustering algorithm (Affymetrix 2007) is a special feature of the Affymetrix® GeneTitan® platform in contrast to a stable cluster file used by Illumina®.

On the Affymetrix® GeneTitan® platform, single-stranded DNA hybridizes to an array probe complementary to one of the flanking sequences of the bi-allelic target variant. Depending on the target variant allele, one of two labelled probes binds to the remaining sequence of the DNA strand resulting in a signal that is interpreted by the software. In the two-colour system of the platform, signal clusters are generated based on signal intensity and signal contrast between the signal types of the samples under consideration (signal A, signal B, or their combination in case of a heterozygous genotype; labelled blue, red or yellow in Figure 4). *A priori* expected cluster positions are then adjusted by the algorithm according to observed hybridization signals to obtain the respective genotype calls. Variants are classified into one of six quality categories based on cluster metrics such as the distance between the signal A and signal B clusters: i) “PolyHighResolution” (PHR) characterized by short distances between signal localization and the respective cluster center, clearly separated clusters, and the appearance of all three signal types mentioned above, ii) “NoMinorHom” comparable with PHR, but without signals for one of the two homozygous signal clusters, iii) “MonoHighResolution” with all genotype signals assigned to only one homozygous signal A or signal B cluster, iv) “CallRateBelowThreshold” comparable to PHR, but with too many missing calls compared to the applied threshold, v) “Other” in case of one or more cluster metrics not passing the thresholds, and vi) “off-target variant” (OTV) describing variants for which a cluster can be observed in addition to the expected signal clusters for homozygous and heterozygous genotype calls (labelled cyan in Figure 4). The latter can arise for example if undetected variants in the flanking regions of the target variant occur, which lead to an unstable hybridization between the array probe and the sample DNA and thus to a reduction of signal intensity. For OTVs, genotype calls with expected signal intensities can be analysed. In addition, the information regarding the presence or absence of putative adjacent variants can be taken into account by distinguishing between samples with reduced and expected signal intensities.

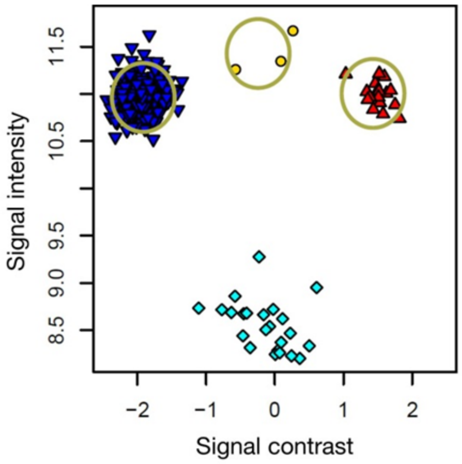

Figure 4: Representative cluster plot for sequence variants categorized as “off-target variant” (OTV). X-axis: contrast of the two colour channels for allele A and B, respectively, Y-axis: signal intensity; blue triangles: homozygous genotype calls for A, yellow circles: heterozygous genotype calls, red triangles: homozygous genotype calls for B, cyan diamonds: OTV genotype calls, yellow circles without colour filling: *a posteriori* genotype cluster positions.

Raw hybridization intensity data processing, clustering, genotype and OTV calling, as well as variant classification according to genotype cluster metrics were performed using Affymetrix® Power Tools version 1.15.0 and the package SNPolisher version 1.3.6.6 (Hong Gao 2012) according to the Axiom® Genotyping Solution Data Analysis Guide. For initial genotype calling, generic *a priori* cluster positions were used since no information about expected cluster positions was available. The three possible genotype clusters were then redefined in *a posteriori* cluster positions, taking the observed genotype call positions into account. Variants were classified based on their respective cluster metrics. In a second, extended analysis different levels of inbreeding were taken into account for *a posteriori* cluster definition because of the high amount of lines in the validation panel that exhibited only a small proportion of heterozygosity in contrast to populations in Hardy-Weinberg equilibrium. The inbred correction was achieved by a parameter that included sample-specific penalties for re-defining *a priori* cluster positions for genotype calling and thus, to adjust the probability of observing a heterozygous call given the inbreeding level of the sample. Based on a range from zero to 16, the following values were assigned: 0 for F1 hybrids, 12 for inbred lines with unclear homozygosity level, and 14 for advanced inbred and doubled haploid lines. Variant classification and results of the analyses with and without inbred correction were compared and a set of randomly selected genotype clusters visually checked. Variants were preferentially selected for building the final 600 k array, if they exhibited stable category assignments with clearly separated clusters to avoid restrictions dependent on the inbreeding level.

For the selection of high-confidence variants for the 600 k array, a voting system was applied using a customized script based on i) their performance on the screening arrays, ii) concordance of array genotyping calls with *in silico* variant calls from sequencing data of the discovery panel, and iii) over- or under-representation of the corresponding physical bin (Unterseer et al. 2014). To ensure a high performance on the final array, the highest weight was assigned to the first criterion focussing on clearly separated genotype clusters with little variance that were not influenced by information regarding the inbreeding level. For the second criterion, the number of calls per variant matching between sequencing and genotyping calls of the lines of the discovery panel was normalized to the total number of calls per variant resulting in a value in the range of zero to one. The criterion of lowest impact, the over- or underrepresentation of a 100 kb bin, was taken into account by calculating the deviation of the number of variants in the corresponding bin to the mean of variants in the five bins up- and downstream, respectively, and scaling the values between minus one and one. The highest scoring sequence variants were selected for the final array. In case of the in total 48,324 SNPs of the 50 k genotyping array (Ganal et al. 2011), which were tiled from both sides on the screening arrays, the probe with the higher rank was included in the final set. If both probes of a variant exhibited the same rank, one probe was chosen randomly. Due to erroneous mapping of 2,669 SNPs of the 50 k array to the B73 reference sequence, a non-polymorphic position was obtained on the screening arrays and these non-validated SNPs were not included on the final array. The top 616,201 variants were selected for the final array design with 45,655 variants originating from the 50 k array.

For OTV validation, sequence reads of four deep sequenced lines were mapped to the B73 reference sequence version 2 (Chia et al. 2012) using the CLC Genomics Workbench version 7.5.1 (http://www.clcbio.com). After standard import of raw sequencing data, the read sequences were trimmed with default parameter setting except that the maximum number of ambiguous nucleotides was set to one. OTVs identified based on the 600 k array were visually checked to investigate sequence variation in the region of the array probe.

2.3 Investigation of population structure and LD

To investigate population structure and LD within datasets A and G, missing genotype calls were imputed based on flanking markers using Beagle version 3.3.1 (Browning and Browning 2009). Analyses of population structure were performed using ADMIXTURE (Alexander et al. 2009). This software was used to estimate the most likely number of groups within a panel of individuals, and the proportion of ancestry per individual that is attributable to one or more of these groups. For a given number of *K* groups, an iterative process was applied to obtain i) an ancestry coefficient matrix based on the fractions of an individual´s genome contributed by each of the *K* groups and ii) the population frequency matrix according to the allele frequencies for each of the *K* groups. Based on the obtained maximum likelihood estimates of the ancestry coefficients and the allele frequencies contributed by each of the *K* groups, the most likely number of groups was evaluated based on a cross-validation procedure. By partitioning the samples into five equally sized folds, prediction errors were estimated by comparing observed, but masked genotypes with the ones predicted by the procedure. As the approach implemented in ADMIXTURE does not account for LD, marker sets were pruned based on an r^2 threshold of 0.8. In this step, a sliding window approach was applied, removing one SNP per pair that exhibited an $r^2 > 0.8$ with a window size of 50 SNPs (sliding by 10%).

Two commonly used measures of LD are r^2 (Hill and Robertson 1968) and *D’* (Lewontin 1964). Both measures are based on the difference between the observed and the expected frequency of the haplotype AB in case of two bi-allelic loci (Lewontin and Kojima 1960):

$$D_{AB} = p_{AB} - p_A p_B$$

with p_{AB} referring to the frequency of haplotypes consisting of the pair of alleles A and B at two loci, and p_A and p_B denoting the frequency of allele A at the first locus and the frequency of allele B at the second locus, respectively. D_{AB} is the coefficient of LD and is also written as *D* in case of two bi-allelic SNPs (Slatkin 2008). The LD measure r^2 is defined as the squared coefficient of LD between the two loci divided by the product of the frequencies of the alleles A, B, and their two alternative alleles, a, and b (Hill and Robertson 1968). It ranges from zero to one, but can only be one in the presence of two haplotypes and equal allele frequencies. Values of r^2 are reduced by mutation and recombination. Lewontin (Lewontin 1964) introduced another measure of LD, *D’*. It is defined as the coefficient of LD divided by the maximum value of *D* given the allele frequencies in the material under study. The largest positive value *D* can take is $p_A p_b$ or $p_a p_B$, whichever is smaller, while the largest negative value *D* can take is either $p_A p_B$ or $p_a p_b$, whichever is smaller. Possible values of D’ range from zero to one like for r^2. In contrast to the latter, the definition of *D*' has the property that the absolute value of *D*' only equals one if at least one of the four possible haplotypes cannot be observed, regardless of the allele frequencies. Thus, values of *D*' that are smaller than one are indicative for historical recombination events.

To obtain an estimate of the level of LD within dataset A, r^2 was calculated between pairs of SNPs as r^2 is considered to be more robust with respect to allele frequencies and sample sizes compared to *D'* (Du et al. 2007). r^2 was calculated between pairs of SNPs on a chromosome within a distance of 50 Mb using PLINK version 1.07 (Purcell et al. 2007). The LD decay distance between SNPs was estimated by plotting the r^2 values of the SNP pairs against their physical distance, fitting a nonlinear regression curve (Hill and Weir 1988) and deriving the crossing point between curve and chosen r^2 threshold of 0.2. LD decay analysis was performed using the R package “synbreed” (Wimmer et al. 2012).

Haplotype blocks refer to a particular combination of adjacent SNPs, which are inherited together, and can be inferred based on a local reduction of diversity (Daly et al. 2001) or recombination (Gabriel et al. 2002; Reich et al. 2001). Haplotype blocks were identified according to the method proposed by Gabriel and colleagues (Gabriel et al. 2002) using the Haploview software (Barrett et al. 2005) via PLINK version 1.90 (Chang et al. 2015). The chosen approach was based on the identification of historical recombination events and thus, on *D'*. As values of *D'* can be inflated in case of small sample sizes or in the presence of rare alleles, confidence intervals for *D'* were constructed based on the observed data for each pair of sites. Pairs of SNPs were defined to be in “strong LD” if the lower boundary of the confidence interval exceeded 0.70 and the upper 0.98, consistent with no historical recombination. Contrary, SNP pairs were considered as revealing “strong evidence for historical recombination” if the upper boundary was below 0.90. A haplotype block was detected based on a contiguous set of SNPs, for which at least 95% of SNP pairs, assigned to either of the two categories, were classified to be in “strong LD”.

2.4 Overview of population genetic metrics

Positive selection can be detected based on deviations of observed allele frequency distributions from expectations under neutrality. For testing neutrality, different estimators of the composite parameter θ are often compared, which is defined by

$$\theta = 4N_e u$$

with N_e denoting the effective population size and u the generation mutation rate (Watterson 1975). Assuming that mutations do not occur at the same position in the sequence more than once and that mutations are segregating in the population unless they get fixed or lost by drift in the absence of selection, θ describes the probability that two alleles, sampled at random from a population, have not been inherited from a recent common ancestor (Hamilton 2009). Three estimators of θ, namely $\hat{\theta}_S$, $\hat{\theta}_\pi$, and $\hat{\theta}_L$, will be introduced below.

Under the infinite-sites model of mutation (Kimura 1969), every mutation can be recognized as a sequence variant that is segregating in the population and thus, θ can be estimated by

$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

using the absolute number of segregating sites *S* in *n* sequences under study (Hamilton 2009; Watterson 1975).

θ can also be estimated based on nucleotide diversity, which is often symbolized as π and known as the average number of pairwise nucleotide differences in *n* sequences (Nei and Li 1979; Tajima 1983). According to Tajima (1983), $\hat{\theta}_\pi$ is defined by

$$\hat{\theta}_\pi = \frac{1}{\binom{n}{2}} \sum_{i}^{n-1} \sum_{j>i}^{n} \pi_{ij}$$

with π_{ij} referring to the number of pairwise nucleotide differences between the *i*th and the *j*th sequence.

$\hat{\theta}_S$ and $\hat{\theta}_\pi$ give comparable results for a random mating population in the absence of selection as proposed by the neutral mutation hypothesis (Kimura 1968, 1969; Kimura 1983). To test this hypothesis, Tajima (Tajima 1989) compared these two estimators of θ:

$$\mathrm{TD} = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{\hat{V}(\hat{\theta}_\pi - \hat{\theta}_S)}}$$

with $\hat{V}(\hat{\theta}_\pi - \hat{\theta}_S)$ denoting the estimated variance of $(\hat{\theta}_\pi - \hat{\theta}_S)$. If TD is close to zero, the neutral mutation hypothesis can explain the observed diversity pattern in the samples under study. Deviations from zero indicate that alternative explanations have to be taken into account. $\hat{\theta}_S$ is affected by the existence of segregating sites, but not by their respective allele frequencies in contrast to $\hat{\theta}_\pi$. Rare alleles for example, are associated with a segregating site, but have limited influence on the average number of pairwise sequence variants. Therefore, negative values of TD indicate an excess of rare alleles associated with directional selection or recent population growth, for example after a bottleneck event. Contrary, positive values indicate an excess of alleles at intermediate frequency as in case of balancing selection or population shrinkage. Thus, TD is sensitive to deviations from neutrality due to demography as well as selection.

To distinguish between demography and selection, $\hat{\theta}_\pi$ can be compared with another estimator of θ, $\hat{\theta}_L$, which includes outgroup information. In case of a selective sweep, hitchhiking due to positive selection is associated with high frequencies of derived alleles. Therefore, the normalized Fay and Wu's *H* statistic (Zeng et al. 2006), *H_{norm}*, contrasts high- and intermediate-frequency variants based on the comparison of $\hat{\theta}_\pi$ and $\hat{\theta}_L$, with

$$H_{norm} = \frac{\hat{\theta}_\pi - \hat{\theta}_L}{\sqrt{\hat{V}(\hat{\theta}_\pi - \hat{\theta}_L)}}$$

and

$$\hat{\theta}_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i\, \zeta_i$$

where ζ_i denotes the number of segregating sites with *i* copies of the derived allele that are observed within *n* samples. If H_{norm} is close to zero, the neutral mutation hypothesis can explain the observed allele frequency pattern. Due to the weighting factor *i*, negative deviations from zero are expected in case of an excess of high derived allele frequencies. In contrast to TD, H_{norm} has been shown to be unaffected by population growth as the overrepresentation of rare alleles is associated with low derived allele frequencies (Zeng et al. 2006). Thus, negative values of H_{norm} are indicative for positive selection.

Selective sweeps can also be detected by comparing the allele frequency distribution of genomic segments to theoretically expected or empirical distributions of allele frequencies. To determine whether a selective sweep occurred at a given set of positions in the genome, the composite likelihood ratio test (CLR) can be performed, as implemented in the software SweepFinder (Nielsen 2005). In the CLR, the likelihoods of two hypotheses are compared: the likelihood of a neutrally evolving sequence that is calculated based on the genome-wide distribution of allele frequencies (null hypothesis), and the likelihood of the alternative hypothesis that a selective sweep gave rise to the observed allele frequency distribution of a genomic region. In case of a sweep, the probability of each individual to escape the sweep is a function of the physical distance between the SNP and the selected site, the effective population size, the recombination rate, and the selection coefficient. Given that some individuals have escaped the selective sweep by recombination, the probability is calculated to observe a specific allele frequency at a given SNP in a certain distance to the selected site under the alternative hypothesis of a sweep. The composite likelihood (CL) of the two hypotheses are then formed by multiplying the probabilities of observing alleles at different frequency classes based on SNPs within the respective genomic region in case of the alternative hypothesis and based on the whole chromosome in case of the null hypothesis. The CLR statistic is given by

$$CLR \;=\; 2 * [\, log(CL_{sweep}) - log(CL_{background})\,]$$

High values of the CLR statistic support the hypothesis that the respective genomic region has been subjected to selection based on an excess of extreme allele frequencies. The test has been shown to be relatively robust under different demographic scenarios and varying recombination rates (Nielsen 2005; Williamson et al. 2007).

Genomic regions under differential selective pressure in two groups are associated with allele frequency differences between these groups, which can be measured by the fixation index F_{ST} (Weir and Cockerham 1984). This index estimates the proportion of genetic variance of allele frequencies between groups in relation to the total genetic variance of their frequencies (Weir and Cockerham 1984). A value of zero indicates comparable allele frequencies in the two groups under study and the absence of population structure if both groups are considered as one group. Contrary, values up to one are suggestive for different allele frequencies in the two groups and thus for population structure in the entire set of individuals.

2.5 Identification of candidate genes in Dent and Flint

For detecting putative signatures of differential selective pressure between Dent and Flint lines of dataset G, a combined approach based on four metrics was applied (Unterseer et al. 2016). Nucleotide diversity π (Tajima 1983) and TD (Tajima 1989) were calculated for each panel of inbred lines using a customized script. The fixation index F_{ST} (Weir and Cockerham 1984) was calculated between the two panels using PLINK version 1.90 (Chang et al. 2015). Metrics were calculated per SNP and averaged over windows of 40 SNPs (sliding by 10%), using the R package “zoo” (Zeileis and Grothendieck 2005). The CLR test was calculated for each panel using the software SweepFinder (Nielsen et al. 2005). For CLR, the grid size was 150 kb, which was the same magnitude as the maximal distance between two SNPs with $r^2 > 0.2$ in dataset A (Unterseer et al. 2014). Windows exhibiting values below the 10% quantile for π and TD as well as above the 90% quantile for F_{ST} and

CLR were submitted for candidate gene analysis based on the B73 reference sequence version 2 (Chia et al. 2012) annotation, version 5b60, which contained 39,656 gene models.

To assess the number of false-positives due to reduced levels of recombination, the lower bounds of recombination events were calculated for Dent and Flint based on dataset G using the four-gamete test, which gives a conservative estimate of recombination events in the history of a sample (Hudson and Kaplan 1985). Values for the pairwise tests of neighbouring SNPs were averaged over 1,000 SNPs and recombination events were reported per Mb. Regions of low recombination rates were defined as regions exhibiting rates below the 10% quantile per chromosome.

2.6 Characterization of candidate genes

Gene ontology terms of the candidate gene sets for Dent and Flint were tested for enrichment using agriGO (Du et al. 2010). Enrichment analyses were performed based on maize gene IDs by applying a hypergeometric test with a Benjamini-Yekutieli correction (Benjamini and Yekutieli 2001) to account for multiple testing and a significance threshold of $\alpha = 0.05$. Pathway analysis was performed using MapMan version 3.5.1 (Thimm et al. 2004) based on the mapping of the first transcript of each gene to the file Zm_B73_5b_FGS_cds_2012.m02 downloaded from the MapMan webpage. Furthermore, candidate genes were assigned to the flowering network in maize based on literature, gene ontology terms, and/or sequence homology to flowering time genes characterized in other species (Unterseer et al. 2016).

Genes identified as candidate genes under differential selective pressure in dataset G were investigated based on additional datasets consisting of 600 k genotyping array data and of whole-genome sequencing data. Gene-wise values of π and F_{ST} were calculated based on dataset L to investigate the contribution of landraces to the reduced candidate gene diversity observed for Dent and Flint based on dataset G. H_{norm} (Zeng et al. 2006) was calculated per gene based on dataset O to test if candidate genes were enriched for high derived allele frequencies compared to non-candidate genes. Gene-wise values of π, TD, and F_{ST} were obtained for dataset S and compared between candidate genes and non-candidate genes to confirm the reduction of candidate gene diversity observed based on dataset G. Gene-wise values of F_{ST} were calculated based on dataset T to investigate allele frequency differences between temperate and tropical lines in case of candidate gene sets and non-candidates using VCFtools version 0.1.11 (Danecek et al. 2011). Analyses based on dataset L and O were performed using customized scripts and analyses based on dataset S using Variscan version 2 (Hutter et al. 2006). For gene-wise calculations, the longest protein-coding transcript, including 5 kb upstream, was used and metrics were calculated if at least five SNPs were available for analysis. For the analysis of exonic, genic, 500 bp, and 5 kb upstream regions, F_{ST} was determined between Dent and Flint lines in case of at least five SNPs per region based on dataset S using VCFtools version 0.1.11 (Danecek et al. 2011). Two-sided Wilcoxon rank sum tests (Wilcoxon 1945) were performed to test for differences between candidate genes and non-candidate genes within pools.

The phenotypic effect of a Flint segment, carrying one or more candidates associated with the flowering network in maize, was investigated based on a Dent–Flint introgression library (Unterseer et al. 2016). Based on two experiments, adjusted means for male and female flowering time were calculated for 97 IL lines of dataset P in a two stage approach using Plabstat (Utz 2011). In the first stage, adjusted entry means were calculated across replicates per location by standard lattice analysis. An outlier detection was performed at the first stage (Anscombe and Tukey 1963) and extreme residuals set to missing (ratio between residual and measurement ≤ -1.9 and ≥ 1.9, respectively). In the second stage, adjusted means of the 97 single-segment IL lines and the Dent parental line were calculated across locations considering genotypes as fixed effects and location and interaction between location and genotype as random effects. Student's *t*-tests were performed to test for significance between adjusted means of flowering time of lines carrying a Flint segment including Flint flowering candidates, Dent flowering candidates and non-candidate genes, respectively. The least significant difference in adjusted means between the 97 single-segment lines and the Dent parent was determined at two significance levels of $\alpha = 0.05$ and $\alpha = 0.05/97$, the latter to correct for multiple testing. To estimate the length of individual donor genome fragments, the distance between markers on the respective donor genome fragment plus half the distance to the adjacent marker flanking the donor genome fragment on either side of the fragment was calculated.

3 Discussion

The objective of the study was the identification of candidate genes under differential selective pressure in two temperate maize germplasm pools. In this chapter, results presented in the two publications underlying this thesis are discussed and complemented by additional findings. The first two sections present the design of the developed genotyping array and its application to resolve population structure and LD extent in the lines under study. The subsequent two sections address the identification of candidate genes and their biological relevance focussing exemplarily on the elucidation of the complex network of flowering time in maize. Potential effects of ascertainment bias and consequences of varying levels of population and LD structure on the sensitivity of the performed selection screen are discussed in section 3.5, followed by a summary of the major findings of the thesis.

3.1 Design of the 600 k genotyping array

High-throughput genotyping has revolutionized genetic analyses in humans, livestock species, crop and model plants in the past decade by offering an efficient alternative to whole genome sequencing for gaining genomic information (Hayes et al. 2013; Langridge and Fleury 2011; Ragoussis 2009). Technological advances in genomic research paved the way for the generation of an increasing number of genotyping arrays for various agronomically important plant species, including rice (Chen et al. 2014), wheat (Winfield et al. 2016), sunflower (Livaja et al. 2016), apple (Bianco et al. 2016), soybean (Wang et al. 2016), brassica (Clarke et al. 2016) and oil palm (Kwong et al. 2016). To ensure the utility of a genotyping array for a wide range of research questions and study designs, its establishment requires the identification of a large number of variants that are polymorphic in a representative discovery panel. For the establishment of a high-density genotyping array for maize, a discovery panel, comprising 30 important founder lines of maize breeding in Europe and the US, was sequenced for variant discovery at intermediate to high coverage (Unterseer et al. 2014). After applying a stringent multi-step filtering procedure, 1.2 M of the 57 M initially identified sequence variants were experimentally investigated by genotyping a validation panel, a diverse panel of 285 maize samples that represented the genetic diversity of European and US temperate maize as well as a set of tropical maize lines. Based on genotype call cluster separation, cluster variance, and cluster position, variants were assigned to one out of six quality categories. As the majority of samples exhibited only a minor level of heterozygosity, category assignment was compared between the assignment with and without inbreeding correction. For inbreeding correction, sample-specific factors were incorporated in the genotype calling algorithm to adjust the probability of observing a heterozygous call given the inbreeding level of the sample. Category assignments changed in 36.2% of all variants upon inbred correction with the category of variants fulfilling all cluster metric criteria and classified as PHR benefitting most. Thus, applying the inbred correction was highly recommended for data analysis using the flexible genotype calling algorithm provided by Affymetrix®. Based on category assignment, physical distribution, and concordance with *in silico* variant calls from sequencing data, a final selection of 609,442 SNPs and 6,759 indels was created for the 600 k genotyping array. Except 262 variants derived from the 50 k array, all remaining variants of the 600 k array (99.9%) were polymorphic in dataset A, for which polymorphism rates of more than 95% were observed for Dent, Flint and F1 hybrids (Unterseer et al. 2014).

The 600 k array included 116,224 variants, which were located in coding regions. Based on the B73 filtered gene set, 26,620 genes (67%) were tagged with at least one variant in their coding, intronic, or untranslated region, compared to 17,520 genes tagged by SNPs of the 50 k array (44%). Including 5 kb up- and downstream regions, 35,089 genes (88%) were represented by at least one variant, thus providing an excellent basis for finding marker-trait associations in targeted and genome-wide approaches. The average distance between two sequence variants was 3.4 kb compared to 45 kb for the 50 k array (Ganal et al. 2011). The physical distribution of the 616 k high-quality variants followed the estimated recombination rate profile along the chromosomes with less sequence variants in centromeric compared to telomeric regions. This reduction of variant numbers around centromeres was also observed in other maize studies (Chia et al. 2012; Gore et al. 2009; Romay et al. 2013) and resulted from the high proportion of repetitive DNA around the centromeres for which no markers could be developed. The high reproducibility of the genotype calls was shown by up to 99.8% of identical genotype calls for replicates and by up to 94.3% of variants with stable Mendelian inheritance in trios consisting of parental lines and the corresponding hybrid. The level of reproducibility of genotype calls was in the same range as reported for the 50 k array (Ganal et al. 2011), thus highlighting the advantage of genotyping arrays to generate robust and reproducible genotype calls. To ensure a high genotype concordance between laboratories and across genotyping platforms in case of the flexible Affymetrix® genotype calling algorithm, a genotype cluster model file was established for stable PHR cluster positions that is available via the Affymetrix® website. Furthermore, 45,655 variants of the

50 k array (Ganal et al. 2011) were successfully validated and included for study comparison.

The 600 k genotyping array is the largest SNP array currently publicly available for maize and represents a powerful tool for fine-mapping of genomic regions, haplotype construction and detection of marker-trait associations. The high-density array was applied to determine population structure and LD extent (Unterseer et al. 2014) and to identify candidate genes under differential selective pressure between Dent and Flint (Unterseer et al. 2016) as shown in the following sections. Furthermore, the array was successfully applied in targeted approaches to narrow down a candidate region for haploid induction (Hu et al. 2016) and to resolve genomic variation underlying expression-based presence/absence variation (Jin et al. 2016). Additional applications of the 600 k array may include its use in the imputation of genotypes from genetic material analysed with lower density marker panels and the saturation of specific genomic regions with SNPs for fine-mapping, map-based cloning studies or marker-assisted selection.

3.2 Population structure and the extent of LD

The identification of population structure is crucial for quantitative genetic or population genetic studies since admixture may affect the estimation of population genetic parameters, the detection of marker-trait associations, or accuracies of genomic prediction. Therefore, dataset A and dataset G were investigated with respect to the population structure underlying 155 temperate and (sub)tropical lines (Unterseer et al. 2014) and the subset of 136 temperate Dent and Flint lines (Unterseer et al. 2016), respectively.

Analysis of population structure revealed seven groups within dataset A and six groups within dataset G, namely the Dent groups BSSS, LSC, Iodent and non-BSSS, the Flint groups Northern and non-Northern Flint as well as an additional group including (sub)tropical lines in case of dataset A. Except for two lines with a presumable contribution of (sub)tropical lines in their pedigree, pool assignment was consistent between dataset A and dataset G for Dent resulting in 14 BSSS, 14 Iodent, nine LSC, and 33 non-BSSS in case of dataset G. This was expected considering that the majority of these lines were US Corn Belt Dent lines and that Corn Belt Dent comprises several heterotic pools, which were established based on few founder lines followed by divergent selection during the last decades (Mikel and Dudley 2006; Nelson et al. 2008; van Heerwaarden et al. 2012). Furthermore, van Heerwaarden and colleagues reported strong genetic differentiation between modern North American maize lines of BSSS, Iodent and non-BSSS, including LSC (van Heerwaarden et al. 2012). In line with this finding, the level of genome-wide differentiation between the four Dent pools based on dataset G was substantial (Table 2). The percentage of polymorphic SNPs observed in pairwise comparisons of BSSS, Iodent and LSC was lower than in pairwise comparisons between one of these pools and non-BSSS (73.5-75.6% vs. 91.6-91.8%). This might result from smaller pool sizes in case of BSSS, Iodent and LSC compared to non-BSSS, but could also indicate heterogeneous selective pressure in the genome.

Table 2: Level of differentiation between the four Dent pools of dataset G. Average values of F_{ST} between BSSS, Iodent, LSC, and non-BSSS are shown in the upper triangle. The respective number of polymorphic SNPs is listed in the lower triangle.

	BSSS	Iodent	LSC	Non-BSSS
BSSS	-	0.299	0.290	0.125
Iodent	409,696	-	0.320	0.170
LSC	413,965	402,595	-	0.144
Non-BSSS	502,539	502,194	501,512	-

Most Flint lines in this thesis were derived from material introduced to Europe around 500 years ago by several expeditions and varying trade routes (Figure 2). The diverse background of the Flint lines under study was reflected by their non-consistent group assignment in dataset A and dataset G. With both datasets, two groups of temperate Flint lines were identified, referred to as Northern and non-Northern Flint according to their geographic distribution within Europe (Unterseer et al. 2014; Unterseer et al. 2016). A total of 24 lines from Germany and France was assigned to Northern Flint in case of both datasets, whereas only 18 of the remaining 42 non-Northern Flint lines of dataset G were also assigned to non-Northern Flint based on dataset A. The remaining non-Northern Flint lines of dataset G were assigned to Northern Flint and a group including (sub)tropical lines based on dataset A (10 and 14 lines, respectively). In line with the absence of a pronounced population structure in the temperate Flint lines under study, a low level of genetic differentiation was observed between the two groups within dataset G (mean F_{ST} of 0.078).

The majority of early European maize hybrids resulted from crosses between US Dent lines and European Flint lines (Barrière et al. 2006). Until the end of the 1970s, the European Flint germplasm was strongly influenced by few founder lines such as F7 and F2, which were derived from the French landrace Lacaune (Barrière et al. 2006). Lines with major Lacaune contribution in their pedigree have been reported to form a distinct genetic group (Camus-

Kulandaivelu et al. 2006) as these lines were derived from the hybridisation zone of the Pyrenees and Galicia, where maize with Caribbean and Northern Flint background mixed (Dubreuil et al. 2006; Mir et al. 2013). The historical importance of maize material from this hybridization zone for the Flint lines under study was seen in the comparison of Flint elite lines and 31 Flint-type European landraces based on dataset L. For most of the landraces from south-western France, the level of F_{ST} between Flint elite lines and Flint-type landraces was lower compared to the average of all 31 Flint-type landraces (Table 3; Figure 5).

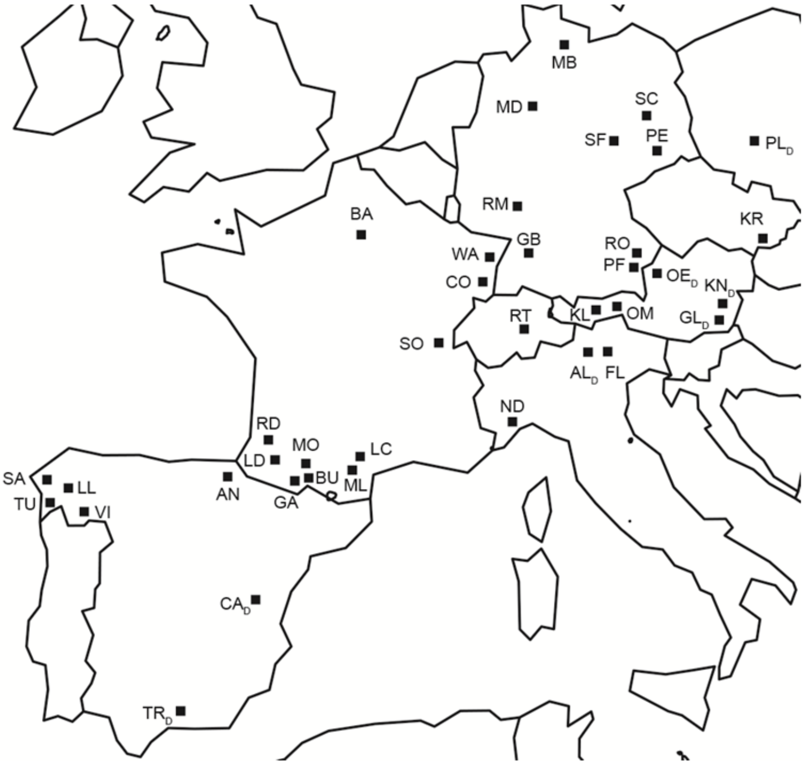

Figure 5: Geographic origin of 38 European landraces included in dataset L. Abbreviations of landraces refer to Table 1. Landraces with Dent-type kernels are indicated by $_D$.

Table 3: Average level of F_{ST} between Flint elite lines and 31 Flint-type landraces of dataset L. Abbreviations of landraces (LR) refer to Table 1.

LR	F_{ST}	LR	F_{ST}	LR	F_{ST}	LR	F_{ST}	LR	F_{ST}
AN	0.097	KL	0.095	ML	0.066	RM	0.134	TU	0.084
BA	0.092	KR	0.085	MO	0.088	RO	0.087	VI	0.078
BU	0.101	LC	0.081	ND	0.153	RT	0.136	WA	0.089
CO	0.073	LD	0.055	OM	0.089	SA	0.078	All	0.094
FL	0.191	LL	0.058	PE	0.101	SC	0.096		
GA	0.058	MB	0.082	PF	0.129	SF	0.082		
GB	0.081	MD	0.137	RD	0.069	SO	0.075		

The extent of LD in a population influences the resolution that can be obtained in genetic analyses. LD decays with physical distance between two sequence variants and therefore, the level of LD within a given set of lines is often described by an average decay distance. This distance can be used as an estimate for SNP densities required in genetic analyses, such as association mapping (Van Inghelandt et al. 2011). LD decay distances of several hundred base pairs up to few kilobase pairs have been reported in highly diverse maize panels (Chia et al. 2012; Lu et al. 2011; Romay et al. 2013; Yan et al. 2009). Most of these panels included tropical material that exhibits a faster LD decay than temperate maize (Lu et al. 2011; Yan et al. 2009). Furthermore, analyses varied with respect to the number and the distribution of markers as well as the applied window sizes (Chia et al. 2012; Riedelsheimer et al. 2012; Romay et al. 2013) and were in some cases restricted to a limited number of genes or loci (Remington et al. 2001; Tenaillon et al. 2001). In this study, the dependency of the LD decay estimation on the chosen window size was exemplarily investigated for chromosome 5 based on the Flint lines of dataset G. The decay distance was slightly underestimated in case of 50 Mb windows compared to a chromosome-wide calculation (174.8 kb vs. 187.1 kb). For smaller window sizes, a clear dependency of the obtained LD decay distance on the chosen window size was observed (Table 4). This dependency resulted from long-range LD that was accounted for when applying larger window sizes. LD over longer distances can arise from population structure, selection, and demographic effects like recurrent bottlenecks (Long et al. 2013; Schaper et al. 2012; Voight et al. 2006). Given the breeding history of maize, long-range LD would not be unexpected and was indeed observed previously. Chia and colleagues reported for example extensive haplotype sharing among improved lines (Chia et al. 2012). Van Heerwaarden and colleagues observed an increase of shared haplotypes among modern inbred lines compared to earlier lines (van Heerwaarden et al. 2012) and Riedelsheimer and colleagues detected considerable long-range LD in elite lines (Riedelsheimer et al. 2012).

Table 4: Dependency of LD decay distance estimates on the chosen window size. Values are shown exemplarily for LD decay calculations based on 66 Flint lines of dataset G in case of chromosome 5.

Window size [kb]	LD decay distance [kb]
5	12.578
50	28.246
500	87.312
5,000	134.205
50,000	174.784
Whole chromosome	187.104

For dataset A, average LD decay distances per chromosome were determined based on LD calculations between pairs of markers within windows of 50 Mb (Unterseer et al. 2014). This window size was chosen as a trade-off between the aim to consider short- as well as long-range LD along the chromosome and computational limitations in the calculation of chromosome-wide LD. LD decayed to an r^2 value of 0.2 within an average distance of 158 kb with smallest distances for (sub)tropical lines (70 kb) and largest for BSSS (36 Mb; Unterseer et al. 2014). LD levels found here were higher compared to previous studies investigating highly diverse maize lines (Chia et al. 2012; Yan et al. 2009). This might be due to the sample panel analysed, which mainly comprised temperate maize lines belonging to distinct germplasm pools and the rather small pool size in BSSS and Iodent. In line with the literature, a substantially higher level of LD was observed in BSSS and Iodent compared to non-BSSS due to a closer relationship and a smaller number of founder lines within BSSS and Iodent compared to non-BSSS (Liu et al. 2003; Mikel and Dudley 2006; Romay et al. 2013).

3.3 Candidate genes for Dent and Flint

A comprehensive investigation of pool-specific targets of selection would be valuable for a better understanding of genomic and phenotypic differences between Dent and Flint and a knowledge-driven optimization of existing breeding schemes. To identify genomic regions under selective pressure in one of the two germplasm groups, dataset G was screened for extreme allele frequencies over extended linked sites in a window-based approach by calculating nucleotide diversity π, Tajima´s D (TD) and the composite likelihood ratio test (CLR) for each of the two germplasm groups (Unterseer et al. 2016). To ensure that the selection signature was specific for one of the two pools, windows had also to be associated with a high level of differentiation between Dent and Flint measured by the fixation index F_{ST}. As changes in allele frequencies were expected to be most prominent in genomic regions under selective pressure compared to the genomic background, an outlier-based approach was applied. Windows were selected for further investigation if they exhibited values below the 10% quantile for π and TD and above the 90% quantile for CLR and F_{ST}. Adjacent windows were combined for candidate gene analysis as the observed changes in allele frequency were likely caused by the same selective sweep event. This resulted in the selection of 265 windows for Dent and 158 windows for Flint with an average size of 331.4 kb and 267.8 kb, respectively (Unterseer et al. 2016). Thus, 4.3% and 2.1% of the maize genome, as calculated from the B73 reference sequence, were identified to be under putative differential selective pressure in Dent and Flint, respectively. Based on these candidate regions, 876 and 545 candidate genes with haplotypes near fixation or fixed in either of the two elite germplasm were identified for Dent and Flint, thus corresponding to 2.2% and 1.4% of the annotated maize gene set, respectively.

The high level of LD in temperate Dent and Flint lines facilitated the detection of selective sweeps, but might have also decreased the power to discriminate between sweep signals caused by genetic hitchhiking due to positive selection and negative background selection in regions with reduced levels of recombination (Charlesworth et al. 1993; Stephan 2010). However, the hypothesis of positive selection being the driving force of the observed allele frequency changes was supported by the observation that with 75% of the Dent and 81% of the Flint candidates the majority of the identified candidate genes were not located in regions with low levels of recombination such as centromeric regions (Unterseer et al. 2016). In addition, candidate genes were enriched for high derived allele frequencies as expected in case of a classic sweep scenario in contrast to background selection. The investigation of dataset S supported the reduced diversity of the identified candidates by a significant reduction of mean gene-wise π and TD in Dent and Flint candidate gene sets compared to non-candidate genes based on whole-genome sequence data. To examine, if genic and upstream regions contributed equally to the differentiation between temperate Dent and Flint, values of F_{ST} were investigated separately for 5 kb and 500 bp upstream regions, genic regions, and exons based on dataset S. F_{ST} values between Dent and Flint lines were significantly higher for candidate gene sets compared to non-candidate genes for all four categories (5 kb, 500 bp, genic, exonic) as expected based on the results obtained from dataset G. However, distributions of F_{ST} values were similar between all four categories in each of the candidate gene sets. Thus, the power to resolve whether selection acted differentially in upstream and genic regions was probably limited by the high level of linkage disequilibrium observed in temperate Dent and Flint (Unterseer et al. 2016). Results of ongoing large-scale whole genome and transcriptome sequencing projects will allow investigating the impact of selection on the regulation of gene activity in these two germplasm pools and their consequence for the differentiation between Dent and Flint.

The assess whether the selection targets were surrounded by long blocks of high LD, haplotype blocks were identified for the Dent and Flint lines of dataset G based on *D'*. A total of 36,085 haplotype blocks was identified for Flint and of 34,250 blocks for Dent with an average length of 39.3 kb and 44.3 kb, respectively. Thus, haplotype blocks were abundant in the genome of both pools, but were significantly longer in Dent compared to Flint (p-value = 3.6e-07). Haplotype blocks that included non-candidate genes, revealed a comparable length of 154.2 kb for Dent and 152.0 kb for Flint on average, though more blocks were found for Dent compared to Flint (Table 5). As expected for sweeps, haplotype blocks that included candidate genes were significantly longer compared to blocks, which did not harbour selection candidates (Table 5). Haplotype blocks were enriched for candidate genes including 59.5% of the Dent and 52.5% of the Flint candidate genes compared to 39.8% and 35.9% of non-candidate genes, respectively. Tracing the extent of haplotype blocks including candidate genes in maize material of different breeding stages will offer an interesting opportunity to increase existing knowledge on how modern Dent and especially Flint germplasm evolved.

Table 5: Characteristics of haplotype blocks identified in Dent and Flint based on dataset G. Total number, mean and median length of blocks including non-candidate and candidate gene sets for Dent (D) and Flint (F), respectively. p-value: significance of difference between the length of haplotype blocks including non-candidate and candidate genes as determined by two-sided Wilcoxon rank sum tests.

Group	Including non-candidate genes			Including candidate genes			p-value
	Number	Mean [kb]	Median[kb]	Number	Mean [kb]	Median[kb]	
D	6,743	154.200	87.620	200	601.800	195.800	<2.2e-16
F	6,522	152.000	85.210	123	371.800	173.600	5.1e-09

As most of the European Flint inbred lines were assumed to be derived from few European landraces (Barrière et al. 2006), the hypothesis was tested that selection on the Flint candidates had occurred prior to modern breeding efforts (Unterseer et al. 2016). The level of differentiation between Flint-type landraces and Flint elite lines was significantly lower for Flint candidates compared to non-candidate genes based on dataset L (F_{ST} of 0.072 vs. 0.095; p-value = 6.0e-04). This finding supported the hypothesis that selection acted on Flint candidates in Flint-type landraces prior to modern line improvement. Lowest levels of F_{ST} were observed for landraces from France, Germany, and Spain, which suggested a major contribution of these Flint-type landraces to the Flint candidate gene diversity observed in the Flint elite lines (Unterseer et al. 2016). This observation was in line with the report of Barrière and colleagues that the German landrace Gelber Badischer Landmais played an important role in the development of flint lines after 1980 (Barrière et al. 2006). The study also reported that lines from Germany and Canada as well as Northern Flint gave rise to significant improvements in early vigour in cooler climates possibly with a contribution of introgressions of maize from the tropical highlands, southern parts of Argentina and Chile. Thus, comparing the genomic composition of these maize groups with European Flint-type landraces might offer additional insights into targets of adaptation to cooler climates. The remaining seven European landraces displayed at least partially Dent-type kernels and their allelic composition was compared to Dent elite lines of dataset G. Dent-type landraces revealed considerable levels of differentiation for Dent candidates compared to non-candidate genes (F_{ST} of 0.164 vs. 0.111, p-value = 0.026), which indicated that European Dent-type landraces exhibited a different allelic composition of the Dent candidates than the Dent elite lines under study.

Considering the phenotypic characteristics of Dent and Flint, candidate gene sets were tested for enrichment of specific biological processes or pathways. No significant GO term enrichment could be observed for the identified genes, though indication for a pool-specific enrichment for genes associated with tetrapyrroles in Dent and with terpenoid metabolism in Flint was observed (Unterseer et al. 2016). Based on sequence similarity to *Arabidopsis thaliana*, Flint candidate genes associated with terpenoid metabolism might be involved in the biosynthesis of *β*-caryophyllenes, which are part of an indirect defence response mechanism against herbivores that has been shown to be largely lost in temperate US Dent in contrast to European Flint (Degen et al. 2004; Kollner et al. 2008; Rasmann et al. 2005). Tetrapyrroles represent precursors of chlorophyll and heme and have been reported to be involved in drought signalling (Nagahatenna et al. 2015). Furthermore, six Flint candidates associated with cold tolerance were identified and for half of these, differential expression upon exposure to chilling temperature has been reported in the literature for maize or the homologous gene in rice. Finally, 30 candidates could be assigned to the flowering network in maize and linked to phenotypic effects as it will be presented in the following section. Bridging the gap between observed genomic differences between Dent and Flint and putative effects of germplasm-specific candidate gene haplotypes on the phenotype is essential for assessing their potential for further improvement of modern maize germplasm. Up to now, RNA expression data across various developmental stages and tissues are mainly available for US Dent lines like B73. Thus, follow-up studies are required for a comprehensive characterization of the identified candidate genes on the transcriptional, structural, and functional level especially in the Flint germplasm.

3.4 Differential selective pressure on the flowering network

Genomic analyses offer an excellent opportunity to gain insights into evolutionary processes. Indications for phenotypic effects of candidate regions from selection screens have been obtained using QTL mapping or genome-wide association studies for example

(Horton et al. 2012; Hufford et al. 2012; Xie et al. 2015). However, only rare examples exist that provide direct support for the functional effects of identified candidate genes or regions in the species under study (Hufford 2016). Flowering time is essential for local adaptation and represents a major determinant for other agronomic traits, such as grain filling and yield. Differences in flowering time have been well described in maize with Dent germplasm flowering on average later compared to Flint (Camus-Kulandaivelu et al. 2006). Thus, candidate genes associated with flowering time were investigated with respect to the differentiation between Dent and Flint in Unterseer et al. (2016). Based on literature, gene ontology terms, and/or sequence homology to flowering time genes characterized in other species, 18 candidate genes could be identified for Dent and 12 candidates for Flint that were associated with the flowering network. For assessing the phenotypic effects of these candidates, dataset P was used. The maize introgression library (IL) included 97 lines, which carried a single Flint genomic segment in a Dent genetic background, and was investigated with respect to changes in flowering time compared to the parental Dent line. Of the 97 lines, 22 lines carried a Flint introgression with one or several of in total 14 Dent and Flint flowering time candidates (six Dent candidate genes and eight Flint candidates). Six of the lines carried a segment with a combination of Dent and Flint candidates. The comparison of flowering time between these 22 lines and the remaining 75 lines, which did not carry a genomic segment with a flowering time candidate identified in the screen, revealed that the seven lines carrying the Flint haplotype of a Flint candidate flowered significantly earlier (93.1 versus 96.1 days, p-value = 0.011). Contrary, nine lines which carried the Flint haplotype of a Dent flowering time candidate did not exhibit a significant shift in flowering time compared to the 75 lines. The obtained results demonstrated that the Flint haplotypes of the Flint flowering candidates promoted earlier flowering in contrast to the respective Dent haplotypes, thus linking candidate genes identified based on allele frequency changes to phenotypic effects. The IL offers a unique study system to investigate the effect of specific haplotypes on the phenotype and a further reduction of the size of the introgressed segments would be highly desirable for further studies. Moreover, the comparison of Dent and Flint can be extended to other candidate genes by phenotyping additional traits, measuring transcriptional and metabolomics data, and by investigating interactions of candidate genes with the genomic background.

Timing of the transition from the vegetative to the reproductive phase is crucial for the adaptation to different environments and the agronomic performance of maize. Many developmental and physiological traits are influenced by flowering time and maturity, which makes the profound understanding of the regulation of the genetic network underlying flowering time highly desirable. The flowering network comprises pathways associated with the integration of environmental signals, e.g. light perception and photoperiod, as well as of endogenous signals via autonomous, age and phytohormone-dependent pathways and has been well characterized for example in rice (Lee and An 2015) and *Arabidopsis thaliana* (Bouche et al. 2016). It has been shown that the progressive adaptation of short-day plants like maize, rice and tomato to temperate climates required the loss of photoperiod sensitivity (Hung et al. 2012; Nakamichi 2015). In maize, the complex genetic architecture of flowering time has been studied in a large number of studies mapping QTL with a meta-analysis revealing 62 flowering time consensus QTL (Chardon et al. 2004). Phenotypic differences in maize flowering time are mainly caused by the accumulation of many small-effect QTL (Buckler et al. 2009). Only a few large-effect genes have been characterized in maize so far (Colasanti et al. 1998; Danilevskaya et al. 2008; Muszynski et al. 2006; Salvi et al. 2002; Vladutu et al. 1999) and were included in a conceptual gene regulatory network model for flowering time control in maize (Dong et al. 2012). In this study, the 30 flowering time candidates were assigned to different pathways within the flowering network based on the function of their homologs in *Arabidopsis thaliana* or maize-specific reports (Unterseer et al. 2016). This revealed that the majority of the Dent candidates was involved in light perception and photoperiod dependent pathways (12 of 18 Dent candidates), whereas in Flint the majority of the candidates was associated with endogenous signal integration and flower developmental processes (10 of 12 Flint candidates). Thus, it could be shown that different pathways of the flowering network were under selective pressure in temperate Dent and Flint. Taking the results of the analysis of the IL lines into account, the Flint-specific haplotypes of Flint candidate genes very likely constitute a promising source for the adaptation of maize germplasm pools to shorter vegetation periods by promoting earlier flowering through endogenous signalling pathways.

The observation that different components of the flowering network were found to be under differential selective pressure in temperate Dent and Flint maize motivated further investigation with specific emphasis on the history of the material under study. As indicated in Figure 2, Corn Belt Dent arose from the historical hybridization between Northern Flint from the north-eastern US and Southern Dent from the south-eastern US roughly 200 years ago with a major contribution from Southern Dent (Anderson and Brown 1952). Therefore, it is likely that the screen for signatures of differential selective pressure in Dent compared to Flint had a high sensitivity with respect to genomic regions tracing back to Southern Dent. As Southern Dents are closely related to southern Mexican varieties with some influence of Caribbean material (Brown and Anderson 1948; Doebley et al. 1988; Liu et al. 2003), genes identified as being under selective pressure in Dent might partially trace back to targets of selection in tropical material. Thus, the following hypothesis could be stated: If genes had been under selective pressure in tropical material, their reduced diversity might have contributed to a reduced diversity in Southern Dent and to the observed low level of diversity of Dent candidates in the Dent lines. To address this hypothesis, the allelic composition of 13 tropical lines was compared to the allelic composition of Dent and Flint lines based on dataset T.

The hypothesis of a relatively close genetic relationship between Dent and tropical maize was supported by a significantly lower level of differentiation between tropical lines and Dent compared to Flint ($F_{ST} = 0.063$ vs. $F_{ST} = 0.114$; p-value < 2.2e-16). If Dent candidate genes experienced selective pressure exclusively in Dent, or if candidate gene haplotypes with different allelic composition were selected in tropical maize and Dent, higher levels of F_{ST} between Dent and tropical maize would be expected for Dent candidates compared to non-candidates. However, no significant change of gene-wise levels of F_{ST} was observed between tropical and Dent lines for Dent candidate genes compared to non-candidate genes (Table 6). Together with the distribution of gene-wise F_{ST} values (Figure 6), this might indicate that the majority of the identified Dent candidates experienced selective pressure in the tropical lines and suggested the presence of targets of selection common to US Dent and tropical maize. Thus, the reduced diversity of Dent candidate genes observed in the Dent lines of dataset G might partially trace back to a reduced diversity of these genes in tropical maize, probably via the contribution of Southern Dent to modern US Corn Belt Dent. For Dent flowering time candidates, gene-wise levels of F_{ST} between tropical and Dent lines were also not significantly different. Five of the 17 genes revealed values of F_{ST} that exceeded 0.1, which were all associated with the response to photoperiod and the circadian system (Unterseer et al. 2016) and might have contributed to the adaptation of Dent to temperate climates. Additional investigation of the genetic composition of these genes in Southern Dent and the effect of their respective haplotypes on the phenotype might provide further support for this hypothesis.

Table 6: Gene-wise F_{ST} values for Dent (D) and Flint (F) candidate gene sets compared to non-candidates based on dataset T. Gene-wise values of F_{ST} between tropical and Dent lines are shown for the comparison of Dent candidates vs. non-candidates and gene-wise levels of F_{ST} between tropical and Flint lines for Flint candidates vs. non-candidates. p-value: significance of difference between non-candidate and candidate genes as determined by two-sided Wilcoxon rank sum tests.

Candidate gene set	**Non-candidate genes**			**Candidate genes**			**p-value**
	Number	**Mean**	**Median**	**Number**	**Mean**	**Median**	
D - All	32,579	0.068	0.048	725	0.079	0.044	0.527
D - Flowering	32,579	0.068	0.048	17	0.067	0.038	0.679
F - All	31,787	0.106	0.082	421	0.185	0.151	< 2.2e-16
F - Flowering	31,787	0.106	0.082	10	0.182	0.152	0.016

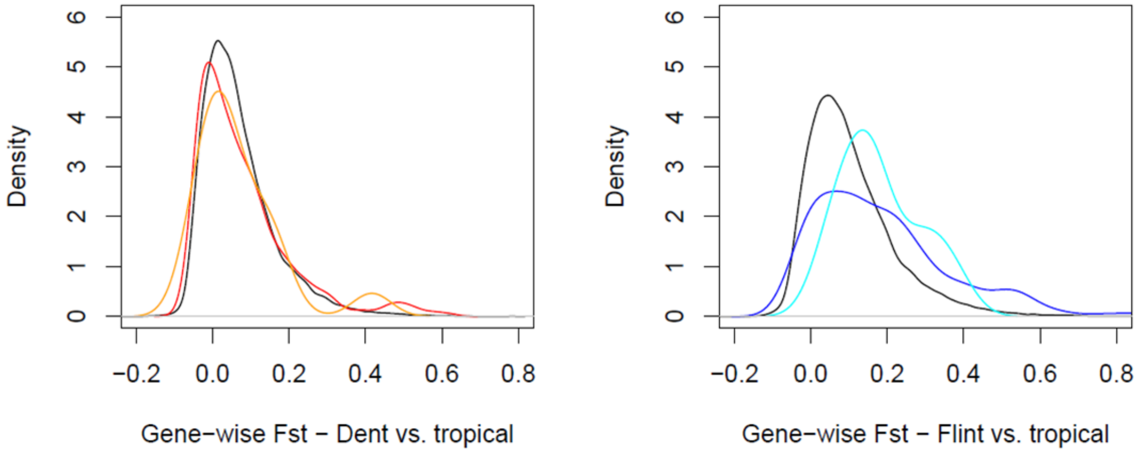

Figure 6: Distribution of gene-wise F_{ST} values between tropical and Dent lines (left) and tropical and Flint lines (right) based on dataset T. Non-candidate genes are shown in black, Dent candidates in red, Dent flowering time candidates in orange, Flint candidates in blue and Flint flowering time candidates in cyan.

The average level of differentiation was significantly higher between tropical and Flint lines than between Dent and Flint lines based on dataset T ($F_{ST} = 0.114$ vs. $F_{ST} = 0.098$; p-value < 2.2e-16). In line with the assumed genetic distance between Flint and tropical maize, Flint candidate genes exhibited significantly higher values of F_{ST} compared to non-candidates (Table 6). This observation might result from selective pressure on these genes in Flint in contrast to tropical maize, but might furthermore point towards differential selection. Differential selective pressure has been reported for example in case of the *Vgt1* locus (Ducrocq et al. 2008), a major QTL for flowering that was also found to be under differential selective pressure in Flint in this study (Unterseer et al. 2016). It regulates a downstream located gene encoding the ethylene-responsive transcription factor *Rap2* (*ZmRap2.7*, *GRMZM2G700665*), which was also identified as a Flint candidate in this study and which exhibited with 0.371 the highest level of gene-wise F_{ST} between Flint and tropical maize within the Flint candidates. The majority of Flint flowering time candidates exhibited high values of gene-wise F_{ST} compared to the average gene-wise F_{ST} for non-candidates (Figure 6). Considering the effect of Flint flowering candidate gene haplotypes on promoting earlier flowering in the IL, the modulation of endogenous signalling might have been of special relevance for the regulation of flowering time and thus, the successful adaptation of Flint to shorter vegetation periods in temperate climates.

3.5 Factors influencing the sensitivity of the selection screen

The sensitivity of the chosen approach for detecting signatures of differential selective pressure seemed to have had a higher sensitivity towards the detection of Dent-specific signatures of selection as indicated by 67.7% more windows and 60.7% more genes identified in Dent compared to Flint based on dataset G. The detection of selection candidates might have been affected by the type of genotyping data as well as by the choice of genetic material as will be discussed in the following.

The SNPs included on the 600 k genotyping array were initially identified based on whole-genome sequence data of the 30 lines of the discovery panel and filtered according to various quality criteria (Unterseer et al. 2014; Unterseer et al. 2016). The sampling bias, which arises during such SNP discovery and selection processes results in a systematic deviation from the theoretically expected allele frequency distribution for a given population due to the non-random sampling of lines for the discovery panel and the sequence variants and is summarized by the term ascertainment bias. Its magnitude is primarily determined by the composition and the size of the discovery panel as the probability to identify rare variants depends on their allele frequencies in the discovery panel (Nielsen et al. 2004). As a consequence, predominantly older variants will be preferentially selected due to their higher allele frequencies and variants included on genotyping arrays are enriched for intermediate frequencies as it was shown for dataset A (Unterseer et al. 2014). Ascertainment bias is not specific to SNP array genotyping data and has been shown for microsatellites (Eriksson and Manica 2011), restriction site polymorphisms (Eller 2001), restriction site associated DNA sequencing data (Arnold et al. 2013), and sequencing data (Pool et al. 2010). Ascertainment bias can affect the estimation of population genetic parameters and LD (Nielsen and Signorovitch 2003) as well as measures of genetic differentiation (Albrechtsen et al. 2010), especially if the degree of bias varies between the material under consideration (McTavish and Hillis 2015). In maize, the effect of ascertainment bias has been shown to be more pronounced in European Flint compared to European Dent for a set of sequence variants included on the 50 k array, which were selected with the aim to detect polymorphisms between two Dent lines, the BSSS line B73 and the LSC line Mo17 (Frascaroli et al. 2013). Several approaches have been proposed for the correction of ascertainment bias (Kuhner et al. 2000; McGill et al. 2013; Nielsen 2000; Wakeley et al. 2001). If sequencing data are available, raw data can be modified to reverse-engineer a particular ascertainment scheme (Albrechtsen et al. 2010). Alternatively, ascertainment bias can be incorporated into theoretical population genetic models if the variant discovery and filtering process is known (Nielsen and Signorovitch 2003). However, it remains challenging to appropriately model the ascertainment scheme underlying the investigated data and to evaluate the modelled sampling distribution of variants and multiple linked loci (Lachance and Tishkoff 2013; Nielsen and Signorovitch 2003).

With the aim to generate a high-density 600 k genotyping array suitable for a broad range of applications, sequence variants were identified based on whole-genome sequence data of a diverse discovery panel. As the sequence reads of the 30 lines were mapped to the B73 Dent reference genome for sequence variant discovery, genomic regions not represented in the B73 reference sequence could not be taken into account (Unterseer et al. 2014). Considering the high genetic variability of the maize genome (Chia et al. 2012; Fu and Dooner 2002; Gore et al. 2009; Lai et al. 2010; Springer et al. 2009), this aspect might be of special relevance for germplasm that is genetically distant from the reference sequence such as Flint or tropical maize. Additionally it has to be considered that sequence reads might not map to the reference sequence in case of diverged genomic regions, which could influence the detection of sequence variants. To enhance the sensitivity of detecting sequence variants in Flint for genomic regions represented in the B73 reference genome, more Flint than Dent lines were sequenced at slightly higher coverage on average (17 lines vs. 13 lines; 18.4- vs. 15.1-fold coverage). Upon quality filtering, selected sets of sequence variants were used to investigate dataset A and dataset G. It is important to note that the discovery panel was representative for the two datasets with respect to the covered geographic area, the contribution of the two germplasm pools, and their allelic composition (Unterseer et al. 2014). Nevertheless, an enrichment of intermediate allele frequencies was observed for dataset A (Unterseer et al. 2014). Window-based values of π were comparable in the selection screen between Dent and Flint (average π of 0.308 vs. 0.310; Unterseer et al. 2016) and were tightly correlated with values of TD (0.905 vs. 0.934). Average values of TD were slightly higher in Dent compared to Flint (0.730 vs. 0.682). This probably resulted from more SNPs being monomorphic in Dent compared to Flint (6.9% vs. 4.6%) and a higher average minor allele frequency of the remaining SNPs in Dent than in Flint (0.241 vs. 0.237). Thus, it can be assumed that in the extent of ascertainment bias was comparable in the two germplasm pools Dent and Flint.

The varying population and LD structure of the Dent and Flint lines of dataset G probably had impact on the sensitivity of the performed selection screen. Haplotype blocks including candidate genes differed in number and size between Dent and Flint with an average length of 601.8 kb and 371.8 kb for Dent and Flint, respectively (Table 5). For Dent, these blocks covered in total 120.4 Mb of the genome in contrast to 45.7 Mb in Flint. In both cases, haplotype blocks might have been even longer, as *D'* has been shown to be underestimated in the presence of ascertainment bias (Nielsen and Signorovitch 2003). The long haplotype blocks including Dent candidates were probably maintained over time and might even trace back to founder lines of modern Dent germplasm. This would be in line with the hypothesis of van Heerwaarden and colleagues that the number of lines contributing to the genetic composition of modern North American germplasm has decreased over time and that the US Dent germplasm was initially derived from a relatively homogeneous landrace population (van Heerwaarden et al. 2012). LD can be maintained by drift in case of small effective population sizes due to a bottleneck event, low levels of recombination, or a combination of these two factors (Hamilton 2009; Hill and Robertson 1968). This might have contributed to the establishment of long blocks with a major haplotype at high frequency in Dent. Furthermore, this likely resulted in the identification of more Dent than Flint candidates as the reduction of diversity and the presence of extreme allele frequencies were extended over more adjacent SNPs in Dent compared to Flint. Therefore, differences in population structure as well as presence of longer haplotype blocks in Dent can be considered as the predominant reason for observing 60.7% more candidates for Dent compared to Flint.

The effect of ascertainment bias on the estimation of the fixation index F_{ST} between elite lines, landraces and tropical lines will be discussed in the following. Landraces are considered as important genetic resources with yet untapped genetic diversity (McCouch et al. 2013). Since landraces were not included in the discovery panel, their level of diversity was probably not fully captured by the 600 k genotyping array. It has been reported that F_{ST} can be affected by ascertainment bias depending on the array design and especially in case of a bias in favour of one of the groups under study (Albrechtsen et al. 2010; McTavish and Hillis 2015). In the 600 k array development, the discovery panel included a diverse panel of Flint lines derived from European landraces. Therefore, Flint elite lines and Flint-type landraces of dataset L can be considered as genetically related material and might share a comparable LD structure. Thus, the comparison of the allelic differentiation between Flint elite lines and Flint-type landraces for Flint candidate genes and non-candidate genes might have been affected by ascertainment bias only to a minor extent. European Dent-type landraces can be assumed to be genetically distinct from the Dent elite lines under study, which were primarily composed of lines derived from US Corn Belt Dent material. In this case, local LD might vary considerably between Dent elite lines and Dent-type landraces and also the amount of genetic diversity captured by the 600 k array. Due to the array design, the SNPs are likely to better reflect the diversity of the Dent elite lines compared to the sequence variation within Dent-type landraces. This probably affected F_{ST} estimates between Dent elite lines and European Dent-type landraces in line with reports of group-biased ascertainment schemes (Albrechtsen et al. 2010; Clark et al. 2005; McTavish and Hillis 2015). For the estimation of F_{ST} between temperate and tropical lines of dataset T, the type of bias was different. Genotype calls of lines obtained from the HapMap2 project (Chia et al. 2012) were combined with genotype calls of the discovery panel if available. Due to the restriction to SNPs with less than 50% missing calls, especially SNPs with a high amount of missing calls in the HapMap2 lines were excluded as the HapMap2 lines were sequenced at lower coverage than the lines of the discovery panel on average (8-fold vs. 12- and 50-fold coverage). Furthermore, polymorphic sites in tropical lines were missed if they were monomorphic in the temperate discovery panel. This probably resulted in the underestimation of the diversity of tropical lines in dataset T. Therefore, no gene-wise diversity statistics were reported for tropical lines. The extent of bias was most likely comparable for F_{ST} estimates between tropical lines and Dent or Flint lines of dataset T, as sites that were polymorphic only in tropical maize were missed in both comparisons.

For SNP array data, off-target variants have been suggested to mitigate effects of ascertainment bias (Didion et al. 2012; Fu et al. 2012). Thus, their potential for genetic analyses was investigated based on dataset A. The detection of off-target variants is conditioned on the target variant, which itself was affected by the filtering steps during array development. Furthermore, the occurrence of off-target variants depends on the genetic distance between the material under study and the B73 reference sequence. As expected for initially undetected variants, the lowest amount of genotype calls with reduced signal intensity was observed for Dent lines, especially of BSSS to which the reference line B73 belongs, and the highest for genetically distant material like Flint and tropical lines based on dataset A. To gain insights into the genetic composition of off-target variants, respective genomic regions were investigated by mapping sequence reads of four deep sequenced lines (three Flint and one Dent line) to the B73 reference sequence. For this analysis, 632 OTVs of dataset A were selected. Reduced signal intensities indicated the presence of off-target variants in case of 1,264 genotype calls. Except eight missing calls, the remaining 1,256 genotype calls revealed expected signal intensities. In case of the latter, the majority of genomic regions could be analysed (760 of 1,256; 61%) and validated (741 of 760; 98%). In case of regions, which were expected to exhibit off-target variants based on array data, sequence mapping information was available for 15% of the regions (192 of 1,264). Investigating those, sequence variants were detected within most flanking regions of the target variant (159 of 192; 83%). The majority of these off target variants were SNPs, but also insertions, deletions or combinations of these types of sequence variation were observed. Thus, off-target variants identified by reduced signal intensity can be attributed to different types of sequence variation. This will probably hamper an incorporation of off-target variants in population genetic analyses as population genetic models usually do not account for the combination of sequence variants arising from different mutational mechanisms with varying mutation rates. However, future studies might address the potential of OTVs to investigate for example structural variation based on differences in signal intensities between samples.

3.6 Conclusion

The focus of this thesis was the identification of genes under differential selective pressure in temperate Dent and Flint. The major conclusions can be summarised as follows:

- A new high density genotyping array, the commercially available Affymetrix® Axiom® Maize Genotyping Array genotyping array, was developed based on sequence data of 30 representative temperate European and US maize lines and validated using a diversity panel. The 616,201 variants included on the array were selected in a multi-step approach to ensure the selection of best quality SNPs and small indels for the analysis of different types of material. As the selected variants have been shown to be polymorphic in a broad maize panel, the 600 k array is well suited for fine-mapping of genomic regions, haplotype construction, and detection of marker-trait associations. It represents the largest currently publically available genotyping array for maize offering an efficient alternative to whole genome sequence data for gaining genomic information in high-throughput and high-density for many studies.

- The investigation of off-target variants was suggested to mitigate ascertainment bias in population genetic analyses in the literature. In case of the 600 k array, a set of variants indicated the presence of additional, initially undetected sequence variants in the flanking regions of the target variant. However, off-target variants were not incorporated in LD and population genetic analyses in this thesis as different types of sequence variation were observed in the flanking regions of the target variants.

- The investigation of a diverse panel of temperate maize lines based on the 600 k array revealed distinct genetic groups with elevated levels of LD. In line with known breeding history, population structure was pronounced in Corn Belt Dent. Contrary, no pronounced population structure was observed in European Flint. The genetic composition of the Flint lines was primarily influenced by the historical introduction of Northern Flint from North America consistent with the literature.

- Hundreds of candidate genes were identified as being under differential selective pressure in temperate Dent and Flint and were corroborated by additional analyses, including the investigation of phenotypic data. Candidate genes were shown to promote early flowering in case of Flint candidate gene haplotypes. Candidate gene analyses indicated that selection acted on germplasm-specific targets within the flowering network. The candidates constitute a promising source of genes for further investigation aiming towards a better understanding of germplasm-specific differences between Dent and Flint at the genomic, transcriptomic and phenomic level.

- Analyses of identified selection candidates were expanded to a large 600 k dataset of 38 European landraces that comprised more than 900 individuals and revealed a major contribution of landraces from Germany, France and Spain to the candidate gene diversity in European Flint lines. It was shown that selective pressure occurred for the majority of Flint candidate genes prior to modern breeding efforts in Flint-type landraces. The generated dataset represents a unique resource that will facilitate a more detailed investigation of landraces and the assessment of their potential to further improve maize breeding in a targeted way.

- Differences in population and LD structure affected the sensitivity of the chosen comparative approach resulting in the identification of more candidate genes for Dent compared to Flint. Considering the history of the material under study, results suggested the presence of partially shared targets of selection between Dent and tropical maize, probably due to a historical contribution of Southern Dent to modern Dent germplasm. Thus, the identified candidates likely contributed not only to the differentiation of temperate Dent and Flint germplasm, but might have partially reflected also the differentiation between Northern Flint and Southern Dent.

4 Summary

Maize provides a rich reservoir of genetic diversity to elucidate the effects of adaptation and selection on the genome. Genotyping arrays represent a powerful tool for characterizing genomic diversity, fine-mapping genomic regions and detecting marker-trait associations. In this thesis, one of the largest publicly available SNP arrays in crop species was developed based on sequencing data of 30 representative temperate maize lines. High-confidence variants were selected and experimentally validated. The Affymetrix® Axiom® Maize Genotyping Array is composed of 616,201 SNPs and small indels that were shown to be polymorphic in a broad genetic diversity panel of worldwide maize, thus ensuring the suitability of the array for a wide range of applications. The potential of the genotyping array to resolve population structure and LD extent in diverse maize germplasm with high resolution was illustrated.

Understanding genomic differences between maize germplasm pools may contribute to a better understanding of the complementarity in heterotic patterns and of mechanisms involved in adaptation to different environments. To elucidate how selection shaped the pool-specific genomic diversity of maize, divergence of two major germplasm pools exploited in maize breeding, Dent and Flint, was investigated on a genome-wide scale. By screening a panel of 136 temperate maize lines for extreme allele frequencies over extended linked sites, candidate genes under differential selective pressure in Dent and Flint were identified. The significant enrichment in derived allele frequencies for these genes provided strong indication that the candidate regions represented selective sweeps. The identified candidates included genes associated with traits that are known to differentiate Dent and Flint like cold tolerance and flowering time. By investigating the effect of the flowering time candidates in a Dent-Flint introgression library, it was shown that the Flint haplotypes of these candidates promoted earlier flowering. Within the flowering network of maize, a Flint-specific enrichment of genes associated with endogenous pathways was discovered in contrast to Dent, where selection seemed to act predominantly on genes involved in the response to environmental factors. Low levels of differentiation of Flint flowering time candidate genes between European Flint elite lines and European landraces indicated a major contribution of landraces from France, Germany, and Spain to the candidate gene diversity of the Flint elite lines. The findings of this study highlight the role of genomic regions that have undergone intense selection and contributed to the differentiation of temperate Dent and Flint. The identification of pool-specific selection signatures enabled insights into the patterns of diversity of temperate Dent and Flint and provides new targets for future functional analyses and crop improvement.

5 Zusammenfassung

Mais bietet auf Grund seiner genetischen Diversität ideale Bedingungen, um die Auswirkungen von Selektion und Adaptation auf das Genom zu erforschen. Hoch-Durchsatz Genotypisierungsarrays sind von zentraler Bedeutung für die umfangreiche Charakterisierung genomischer Diversität, eine genauere Kartierung von Genen und die verbesserte Detektion von genetischen Markern, die mit phänotypischen Merkmalsausprägungen in Zusammenhang stehen. Im Rahmen dieser Dissertation wurde einer der umfangreichsten, kommerziell erwerbbaren Arrays für Kulturpflanzen entwickelt. Basierend auf den Sequenzdaten von 30 repräsentativen Maislinien aus den gemäßigten Breiten wurden hoch-qualitative Sequenzvarianten ausgewählt und experimentell validiert. Der Affymetrix® Axiom® Maize Genotyping Array erfasst 616.201 Genompositionen, die in einer Vielzahl von Maislinien variabel sind und dadurch seine Eignung für zahlreiche Anwendungen gewährleisten. Mit Hilfe des Arrays wurde die Populationsstruktur einer Auswahl von Maislinien aus dem Dent- und Flintpool mit großer Genauigkeit erfasst und das Ausmaß des Kopplungsungleichgewichts innerhalb der Gruppen geschätzt.

Um die molekularen Grundlagen der Heterosis sowie lokaler Adaptation zu verstehen, müssen zunächst die genomischen Unterschiede zwischen Maisgruppen erfasst werden. In der vorliegenden Arbeit wurden an Hand von lokalen Änderungen der Allelfrequenzen genomweit Gene identifiziert, die spezifisch in Dent oder Flint unter Selektionsdruck standen. Die Anreicherung evolutionär junger Allele von hoher Frequenz in den Kandidatengenen untermauerte die Hypothese, dass primär positiv selektierte Gene identifiziert worden waren. Zahlreiche Kandidatengene konnten mit Merkmalen in Verbindung gebracht werden, deren Ausprägung sich zwischen Dent und Flint unterscheiden, wie beispielsweise Kühletoleranz und Blühzeitpunkt. In Bezug auf die Regulation des Blühzeitpunktes schienen Kandidaten für Dent überwiegend die Integration externer Signale zu modulieren, wohingegen ein Großteil der Kandidaten für Flint endogene Signalwege beeinflusste. Hierbei wurde mittels einer Introgressionsbibliothek gezeigt, dass die Flinthaplotypen der Kandidaten einen positiven Effekt auf einen frühen Blühzeitpunkt hatten. Zudem wurde gezeigt, dass ein Großteil der Flintkandidaten bereits in europäischen Landrassen unter Selektionsdruck stand und insbesondere Landrassen aus Frankreich, Deutschland und Spanien den europäischen Flintpool maßgeblich prägten. Die gewonnenen Erkenntnisse tragen entscheidend zu einer umfassenden Charakterisierung von Dent und Flint bei und lieferten zahlreiche Kandidaten für künftige funktionale Studien und eine gezielte genetische Verbesserung von modernem Zuchtmaterial.

6 References

Affymetrix I (2007) BRLMM-P: a genotype calling method for the SNP 5.0 array.

Ahn S, Tanksley SD (1993) Comparative linkage maps of the rice and maize genomes. Proceedings of the National Academy of Sciences of the United States of America 90:7980-7984

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27:2534-2547

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655-1664

Anderson E, Brown WL (eds) (1952) Origin of Corn Belt maize and its genetic significance. Iowa State Univ. Press, Ames, Iowa

Anscombe FJ, Tukey JW (1963) The examination and analysis of residuals. Technometrics 5:141-160

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol 22:3179-3190

Baranwal VK, Mikkilineni V, Zehr UB, Tyagi AK, Kapoor S (2012) Heterosis: emerging ideas about hybrid vigour. J Exp Bot 63:6309-6314

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265

Barrière Y, Alber D, Dolstra O, Lapierre C, Motto M, et al. (2006) Past and prospects of forage maize breeding in Europe. II. History, germplasm evolution and correlative agronomic changes. Maydica 51:435-449

Bauer E, Falque M, Walter H, Bauland C, Camisan C, et al. (2013) Intraspecific variation of recombination rate in maize. Genome Biol 14:R103

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13:969-980

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? Trends Ecol Evol 20:435-440

Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, et al. (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193:1073-1081

Beissinger TM, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, et al. (2014) A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. Genetics 196:829-840

Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, et al. (2016) Recent demography drives changes in linked selection across the maize genome. Nat Plants 2:16084

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Statist 29:1165-1188

Bennetzen JL, Hake S (2009) Handbook of maize : genetics and genomics. Springer, New York

Bianco L, Cestaro A, Linsmith G, Muranty H, Denance C, et al. (2016) Development and validation of the Axiom Apple480K SNP genotyping array. Plant J 86:62-74

Birchler JA, Yao H, Chudalayandi S, Vaiman D, Veitia RA (2010) Heterosis. Plant Cell 22:2105-2112

Bouche F, Lobet G, Tocquin P, Perilleux C (2016) FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. Nucleic Acids Res 44:D1167-1171

Bouchet S, Servin B, Bertin P, Madur D, Combes V, et al. (2013) Adaptation of maize to temperate climates: mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the *Vgt2* (*ZCN8*) locus. Plos One 8:e71377

Brown WL, Anderson E (1947) The Northern flint corns. Ann Mo Bot Gard 34:1-22

Brown WL, Anderson E (1948) The Southern Dent corns. Ann Mo Bot Gard 35:255-268

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210-223

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, et al. (2009) The genetic architecture of maize flowering time. Science 325:714-718

Bukowski R, Guo X, Lu Y, Zou C, He B, et al. (2015) Construction of the third generation *Zea mays* haplotype map. bioRxiv preprint http://dx.doi.org/10.1101/026963

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, et al. (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *dwarf8* gene. Genetics 172:2449-2463

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, et al. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7

Chardon F, Virlon B, Moreau L, Falque M, Joets J, et al. (2004) Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. Genetics 168:2169-2185

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303

Chen H, Xie W, He H, Yu H, Chen W, et al. (2014) A high-density SNP genotyping array for rice biology and molecular breeding. Mol Plant 7:541-553

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44:803-807

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496-1502

Clarke WE, Higgins EE, Plieske J, Wieseke R, Sidebottom C, et al. (2016) A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. Theor Appl Genet

Colasanti J, Yuan Z, Sundaresan V (1998) The *indeterminate* gene encodes a zinc finger protein and regulates a leaf-generated signal required for the transition to flowering in maize. Cell 93:593-603

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-232

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156-2158

Danilevskaya ON, Meng X, Hou Z, Ananiev EV, Simmons CR (2008) A genomic and expression compendium of the expanded PEBP gene family from maize. Plant Physiol 146:250-264

Degen T, Dillmann C, Marion-Poll F, Turlings TC (2004) High genetic variability of herbivore-induced volatile emission within a broad range of maize inbred lines. Plant Physiol 135:1928-1938

Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, et al. (2012) Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics 13:34

Doebley J, Wendel JD, Smith JSC, Stuber CW, Goodman MM (1988) The origin of Cornbelt maize - the isozyme evidence. Econ Bot 42:120-131

Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, et al. (2012) A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. Plos One 7:e43450

Du C, Swigonova Z, Messing J (2006) Retrotranspositions in orthologous regions of closely related grass species. BMC Evol Biol 6:62

Du FX, Clutter AC, Lohuis MM (2007) Characterizing linkage disequilibrium in pig populations. Int J Biol Sci 3:166-178

Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res 38:W64-70

Dubreuil P, Warburton M, Chastanet M, Hoisington D, Charcosset A (2006) More on the introduction of temperate maize into Europe: large-scale bulk SSR genotyping and new historical elements. Maydica 51:281-291

Ducrocq S, Madur D, Veyrieras JB, Camus-Kulandaivelu L, Kloiber-Maitz M, et al. (2008) Key impact of *Vgt1* on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178:2433-2437

Durand E, Tenaillon MI, Raffoux X, Thepot S, Falque M, et al. (2015) Dearth of polymorphism associated with a sustained response to selection for flowering time in maize. BMC Evol Biol 15:103

Eller E (2001) Effects of ascertainment bias on recovering human demographic history. Hum Biol 73:411-427

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Plos One 6:e19379

Eriksson A, Manica A (2011) Detecting and removing ascertainment bias in microsatellites from the HGDP-CEPH panel. G3 (Bethesda) 1:479-488

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193:929-941

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413

Feng SQ, Chen XL, Wu SJ, Chen XS (2015) Recent advances in understanding plant heterosis. Agric Sci 6:1033-1038

Frascaroli E, Schrag TA, Melchinger AE (2013) Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. Theor Appl Genet 126:133-141

Fu C-P, Welsh CE, P.-M. dVF, McMillan L (2012) Inferring ancestry in admixed populations using microarray probe intensities. Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine - BCB. ACM Press, New York, pp 105–112

Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. P Natl Acad Sci USA 99:9573-9578

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229

Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, et al. (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. Plos One 6:e28334

Giraud H, Lehermeier C, Bauer E, Falque M, Segura V, et al. (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. Genetics 198:1717-1734

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, et al. (2009) A first-generation haplotype map of maize. Science 326:1115-1117

Hamilton MB (2009) Population genetics. Wiley-Blackwell, Chichester, UK ; Hoboken, NJ

Hayes BJ, Lewin HA, Goddard ME (2013) The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends Genet 29:206-214

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335-2352

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226-231

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

Hirsch CN, Flint-Garcia SA, Beissinger TM, Eichten SR, Deshpande S, et al. (2014a) Insights into the effects of long-term artificial selection on seed size in maize. Genetics 198:409-421

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, et al. (2014b) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121-135

Hong Gao TW, Ali Pirani, Yiping Zhan, Yontao Lu and Mei-mei Shen (2012) SNPolisher: Tools for SNP Classification, Visualization and OTV Genotyping (R package version 1.3.6.6)

Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, et al. (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet 44:212-216

Hu H, Schrag TA, Peis R, Unterseer S, Schipprack W, et al. (2016) The genetic basis of haploid induction in maize identified with a novel genome-wide association method. Genetics 202:1267-1276

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147-164

Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, et al. (2012) Comparative population genomics of maize domestication and improvement. Nat Genet 44:808-811

Hufford MB (2016) Comparative genomics provides insight into maize adaptation in temperate regions. Genome Biol 17:155

Hung HY, Shannon LM, Tian F, Bradbury PJ, Chen C, et al. (2012) *ZmCCT* and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. P Natl Acad Sci USA 109:E1913-1921

Hutter S, Vilella AJ, Rozas J (2006) Genome-wide DNA polymorphism analyses using VariScan. Bmc Bioinformatics 7

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. P Natl Acad Sci USA 101:10667-10672

Jiao YP, Zhao HN, Ren LH, Song WB, Zeng B, et al. (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44:812-U124

Jin M, Liu H, He C, Fu J, Xiao Y, et al. (2016) Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Sci Rep 6:18936

Kaeppler S (2012) Heterosis: many genes, many mechanisms - end the search for an undiscovered unifying theory. ISRN Botany 682824

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513-1524

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624-626

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893-903

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge Cambridgeshire ; New York

Kollner TG, Held M, Lenk C, Hiltpold I, Turlings TC, et al. (2008) A maize (E)-beta-caryophyllene synthase implicated in indirect defense responses against herbivores is not expressed in most American maize varieties. Plant Cell 20:482-494

Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics 156:439-447

Kwong QB, Teh CK, Ong AL, Heng HY, Lee HL, et al. (2016) Development and validation of a high density SNP genotyping array for african oil palm. Mol Plant

Lachance J, Tishkoff SA (2013) SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays 35:780-786

Lai J, Li R, Xu X, Jin W, Xu M, et al. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027-1030

Langridge P, Fleury D (2011) Making the most of 'omics' for crop breeding. Trends Biotechnol 29:33-40

Lee Y-S, An G (2015) Complex regulatory networks of flowering time in rice. J Rice Res 3:141

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458-472

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49-67

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175-195

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079

Liu K, Goodman M, Muse S, Smith JS, Buckler E, et al. (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117-2128

Livaja M, Unterseer S, Erath W, Lehermeier C, Wieseke R, et al. (2016) Diversity analysis and genomic prediction of *Sclerotinia* resistance in sunflower using a new 25 K SNP genotyping array. Theor Appl Genet 129:317-329

Lobell DB, Tebaldi C (2014) Getting caught with our plants down: the risks of a global crop yield slowdown from climate trends in the next two decades. Environ Res Lett 9

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, et al. (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet 45:884-890

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, et al. (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun 6:6914

Lu Y, Shah T, Hao Z, Taba S, Zhang S, et al. (2011) Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. Plos One 6:e24861

Maruyama T, Fuerst PA (1984) Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. Genetics 108:745-763

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, et al. (2002) A single domestication for maize shown by multilocus microsatellite genotyping. P Natl Acad Sci USA 99:6080-6084

McCouch S, Baute GJ, Bradeen J, Bramel P, Bretting PK, et al. (2013) Agriculture: Feeding the future. Nature 499:23-24

McGill JR, Walkup EA, Kuhner MK (2013) Correcting coalescent analyses for panel-based SNP ascertainment. Genetics 193:1185-1196

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303

McTavish EJ, Hillis DM (2015) How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics 16:266

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28:659-669

Mikel MA, Dudley JW (2006) Evolution of North American dent corn from public to proprietary germplasm. Crop Sci 46:1193-1205

Millet EJ, Welcker C, Kruijer W, Negro S, Coupel-Ledru A, et al. (2016) Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios. Plant Physiol 172:749-764

Mir C, Zerjal T, Combes V, Dumas F, Madur D, et al. (2013) Out of America: tracing the genetic footprints of the global diffusion of maize. Theor Appl Genet 126:2671-2682

Morgante M, De Paoli E, Radovic S (2007) Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10:149-155

Muirhead CA (2001) Consequences of population structure on genes under balancing selection. Evolution 55:1532-1541

Muszynski MG, Dam T, Li B, Shirbroun DM, Hou Z, et al. (2006) *Delayed flowering1* encodes a basic leucine zipper protein that mediates floral inductive signals at the shoot apex in maize. Plant Physiol 142:1523-1536

Nagahatenna DS, Langridge P, Whitford R (2015) Tetrapyrrole-based drought stress signalling. Plant Biotechnol J 13:447-459

Nakamichi N (2015) Adaptation to the local environment by modifications of the photoperiod response in crops. Plant Cell Physiol 56:594-604

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

Nelson PT, Coles ND, Holland, J. B., Bubeck DM, Smith S, et al. (2008) Molecular characterization of maize inbreds with expired U.S. plant variety protection. Crop Sci 48:1673-1685

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931-942

Nielsen R, Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor Popul Biol 63:245-255

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168:2373-2382

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197-218

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15:1566-1575

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19:826-837

Piperno DR, Holst I, Winter K, McMillan O (2014) Teosinte before domestication: experimental study of growth and phenotypic variability in late Pleistocene and early Holocene environments Quatern Int 363:65-77

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. Genome Res 20:291-300

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evolution 59:2312-2323

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575

Qanbari S, Gianola D, Hayes B, Schenkel F, Miller S, et al. (2011) Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics 12:318

Ragoussis J (2009) Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet 10:117-133

Rasmann S, Kollner TG, Degenhardt J, Hiltpold I, Toepfer S, et al. (2005) Recruitment of entomopathogenic nematodes by insect-damaged maize roots. Nature 434:732-737

Rebourg C, Chastanet M, Gouesnard B, Welcker C, Dubreuil P, et al. (2003) Maize introduction into Europe: the history reviewed in the light of molecular data. Theor Appl Genet 106:895-903

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. (2001) Linkage disequilibrium in the human genome. Nature 411:199-204

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. P Natl Acad Sci USA 98:11479-11484

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, et al. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217-220

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, et al. (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14:R55

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837

Salvi S, Tuberosa R, Chiapparino E, Maccaferri M, Veillet S, et al. (2002) Toward positional cloning of *Vgt1*, a QTL controlling the transition from the vegetative to the reproductive phase in maize. Plant Mol Biol 48:601-613

Schaper E, Eriksson A, Rafajlovic M, Sagitov S, Mehlig B (2012) Linkage disequilibrium under recurrent bottlenecks. Genetics 190:217-229

Schmidt W (2003) Hybridzüchtung bei der KWS SAAT AG. 54 Tagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs, Gumpenstein

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences of the United States of America 108:4069-4074

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115

Schnell FW (1992) Maiszüchtung und die Züchtungsforschung in der Bundesrepublik Deutschland. Vorträge Pflanzenzüchtung 22:27-44

Schön CC, Dhillon BS, Utz HF, Melchinger AE (2010) High congruency of QTL positions for heterosis of grain yield in three crosses of maize. Theor Appl Genet 120:321-332

Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. Nat Rev Genet 16:727-740

Schrider DR, Mendes FK, Hahn MW, Kern AD (2015) Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200:267-284

Sekhon RS, Hirsch CN, Childs KL, Breitzman MW, Kell P, et al. (2014) Phenotypic and transcriptional analysis of divergently selected maize populations reveals the role of developmental timing in seed size determination. Plant Physiol 165:658-669

Shull G (1908) The composition of a field of maize. Rep Am Breeders Assoc 4:296–301

Shull GH (1909) A pure-line method of corn breeding. American Breeders Association, Cold Spring Harbor, N. Y.

Slatkin M (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477-485

Smith CW, Betrán J, Runge ECA (2004) Corn : origin, history, technology, and production. John Wiley, Hoboken, N.J.

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5:e1000734

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. Philos Trans R Soc Lond B Biol Sci 365:1245-1253

Swanson-Wagner R, Briskine R, Schaefer R, Hufford MB, Ross-Ibarra J, et al. (2012) Reshaping of the maize transcriptome by domestication. Proc Natl Acad Sci U S A 109:11878-11883

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, et al. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20:1689-1699

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437-460

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Teixeira JE, Weldekidan T, de Leon N, Flint-Garcia S, Holland JB, et al. (2015) Hallauer's Tuson: a decade of selection for tropical-to-temperate phenological adaptation in maize. Heredity (Edinb) 114:229-240

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, et al. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). P Natl Acad Sci USA 98:9161-9166

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, et al. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914-939

Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, et al. (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15:823

Unterseer S, Pophaly SD, Peis R, Westermeier P, Mayer M, et al. (2016) A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. Genome Biol 17:137

Utz HF (2011) PLABSTAT - A computer program for statistical analysis of plant breeding experiments. Version 3A, Universität Hohenheim, Germany

van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, et al. (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. P Natl Acad Sci USA 108:1088-1092

van Heerwaarden J, Hufford MB, Ross-Ibarra J (2012) Historical genomics of North American maize. P Natl Acad Sci USA 109:12420-12425

Van Inghelandt D, Reif JC, Dhillon BS, Flament P, Melchinger AE (2011) Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. Theor Appl Genet 123:11-20

Vann L, Kono T, Pyhajarvi T, Hufford MB, Ross-Ibarra J (2015) Natural variation in teosinte at the domestication locus *Teosinte branched1* (*Tb1*). PeerJ 3:e900

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, et al. (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. P Natl Acad Sci USA 99:9650-9655

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. Annu Rev Genet 47:97-120

Vladutu C, McLaughlin J, Phillips RL (1999) Fine mapping and characterization of linked quantitative trait loci involved in the transition of the maize apical meristem from vegetative to generative structures. Genetics 153:993-1007

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. Plos Biol 4:e72

Voss-Fels K, Snowdon RJ (2016) Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotechnol J 14:1086-1094

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001) The discovery of single-nucleotide polymorphisms - and inferences about human demographic history. Am J Hum Genet 69:1332-1347

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, et al. (2005) The origin of the naked grains of maize. Nature 436:714-719

Wang J, Chu S, Zhang H, Zhu Y, Cheng H, et al. (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6:20728

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256-276

Wei F, Zhang J, Zhou S, He R, Schaeffer M, et al. (2009) The physical and genetic framework of the maize B73 genome. PLoS Genet 5:e1000715

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bull 1:80-83

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3:e90

Wills DM, Whipple CJ, Takuno S, Kursel LE, Shannon LM, et al. (2013) From many, one: genetic control of prolificacy during maize domestication. PLoS Genet 9:e1003604

Wimmer V, Albrecht T, Auinger HJ, Schon CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086-2087

Winfield MO, Allen AM, Burridge AJ, Barker GL, Benbow HR, et al. (2016) High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. Plant Biotechnol J 14:1195-1206

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, et al. (2005) The effects of artificial selection of the maize genome. Science 308:1310-1314

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

Wu Q, Zheng P, Hu Y, Wei F (2014) Genome-scale analysis of demographic history and adaptive selection. Protein Cell 5:99-112

Xie W, Wang G, Yuan M, Yao W, Lyu K, et al. (2015) Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. Proc Natl Acad Sci U S A 112:E5411-5419

Xu S, Yang Z, Zhang E, Jiang Y, Pan L, et al. (2014) Nucleotide diversity of maize *ZmBT1* gene and association with starch physicochemical properties. Plos One 9:e103627

Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, et al. (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17:2859-2872

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

Zeileis A, Grothendieck G (2005) zoo: S3 infrastructure for regular and irregular time series. J Stat Softw 14

Zeng K, Fu YX, Shi SH, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174:1431-1439

7 Publications

This appendix contains reprints of the two publications underlying this thesis. The two publications including supporting material can be accessed via the following links:

Unterseer et al. (2014)
http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-823

Unterseer et al. (2016)
http://www.genomebiology.com/2016/17/1/137

Unterseer *et al. BMC Genomics* 2014, **15**:823
http://www.biomedcentral.com/1471-2164/15/823

RESEARCH ARTICLE Open Access

A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array

Sandra Unterseer[1], Eva Bauer[1*], Georg Haberer[2], Michael Seidel[2], Carsten Knaak[3], Milena Ouzunova[3], Thomas Meitinger[4], Tim M Strom[4], Ruedi Fries[5], Hubert Pausch[5], Christofer Bertani[6], Alessandro Davassi[6], Klaus FX Mayer[2] and Chris-Carolin Schön[1*]

Abstract

Background: High density genotyping data are indispensable for genomic analyses of complex traits in animal and crop species. Maize is one of the most important crop plants worldwide, however a high density SNP genotyping array for analysis of its large and highly dynamic genome was not available so far.

Results: We developed a high density maize SNP array composed of 616,201 variants (SNPs and small indels). Initially, 57 M variants were discovered by sequencing 30 representative temperate maize lines and then stringently filtered for sequence quality scores and predicted conversion performance on the array resulting in the selection of 1.2 M polymorphic variants assayed on two screening arrays. To identify high-confidence variants, 285 DNA samples from a broad genetic diversity panel of worldwide maize lines including the samples used for sequencing, important founder lines for European maize breeding, hybrids, and proprietary samples with European, US, semi-tropical, and tropical origin were used for experimental validation. We selected 616 k variants according to their performance during validation, support of genotype calls through sequencing data, and physical distribution for further analysis and for the design of the commercially available Affymetrix® Axiom® Maize Genotyping Array. This array is composed of 609,442 SNPs and 6,759 indels. Among these are 116,224 variants in coding regions and 45,655 SNPs of the Illumina® MaizeSNP50 BeadChip for study comparison. In a subset of 45,974 variants, apart from the target SNP additional off-target variants are detected, which show only a minor bias towards intermediate allele frequencies. We performed principal coordinate and admixture analyses to determine the ability of the array to detect and resolve population structure and investigated the extent of LD within a worldwide validation panel.

Conclusions: The high density Affymetrix® Axiom® Maize Genotyping Array is optimized for European and American temperate maize and was developed based on a diverse sample panel by applying stringent quality filter criteria to ensure its suitability for a broad range of applications. With 600 k variants it is the largest currently publically available genotyping array in crop species.

Keywords: High density genotyping array, Maize, SNP

* Correspondence: e.bauer@tum.de; chris.schoen@tum.de
[1]Plant Breeding, Centre of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising, Germany
Full list of author information is available at the end of the article

Background

High-throughput genotyping has revolutionized genetic analyses in humans, livestock species, crop and model plants in the past decade [1-3]. Covering genomes with high resolution, single nucleotide polymorphism (SNP) genotyping arrays facilitate the detection of associations between SNPs and phenotypes. They represent a powerful tool for dissecting complex traits via genome-wide association studies (GWAS) or quantitative trait locus (QTL) analysis as well as for fine mapping genes of interest and forward genetics cloning strategies [4-7]. In addition, they are broadly used in crop and livestock breeding for germplasm characterization and marker assisted selection [8]. The availability of high density genotyping arrays has enabled breakthroughs in genome-wide approaches such as genomic prediction and detection of selection signatures [9-12]. Here, we describe the development of the currently largest publicly available SNP array in crop species and discuss its potential for different applications in maize.

Maize is one of the most important crops worldwide serving as food, livestock feed, and component of industrial products. A key step in corn production was the establishment of divergent heterotic patterns for hybrid breeding [13]. Most worldwide hybrid breeding programs exploit heterotic effects between different subgroups within the Dent pool, whereas crosses between the two maize pools, Dent and Flint, are mainly used in hybrid breeding for the cooler regions in Central Europe. Maize production has continuously risen over time, but to further increase selection gain and accelerate breeding processes profound knowledge is required regarding genes and genomic regions involved in agronomically important traits.

Genotyping arrays offer an efficient alternative to whole genome sequence data for gaining genomic information in high-throughput. However, the establishment of a high density genotyping array requires the identification of a large number of variants polymorphic in a representative discovery panel to ensure its utility for a wide range of approaches and study designs. In maize, the identification of sequence variants for genomic analyses faces specific challenges due to its evolutionary history and high variability of its genome. As an ancient polyploid species, the maize genome is characterized by numerous duplicated chromosomal regions giving rise to paralogous sequences [14-16]. A reference sequence exists for maize, which covers around 90% of the 2.4 Gb genome of inbred line B73 (AGP_v2), but the high amount of transposable elements, paralogs, copy number variants (CNV) as well as structural variants like presence/absence variants (PAV), is a challenge for reliable sequence read alignment and variant identification due to ambiguous sequence read mapping results [15,17,18]. Despite recent reports like the comprehensive genotyping of the USA national maize inbred seed bank [19] using SNPs identified through genotyping by sequencing (GBS) at low sequence coverage [20], sequencing-based approaches such as GBS have to cope with large amounts of missing data and require the establishment of demanding bioinformatics pipelines and imputing algorithms, which may not be routine in all labs.

The highest resolution of a commercially available genotyping array for maize has been achieved by the Illumina® MaizeSNP50 BeadChip [21]. It has been used extensively for genetic studies [22-25] and is composed of 50 k usable SNPs. This number of SNPs is in the same range as for recently published genotyping arrays for rice [8], soybean [26], and wheat [27], but much lower compared to high density genotyping arrays which are available for animal species, e.g. chicken [28] and cattle with 648 k and 777 k, respectively [29,30], as well as for humans with more than 900 k SNP variants [5]. Especially for maize with its large genome size and high level of diversity, high marker resolution is desirable. In addition, linkage disequilibrium (LD) decays rapidly in some germplasm, e.g. in landraces or highly diverse sample panels [31] emphazising the requirement of higher marker densities than so far available on genotyping arrays.

We selected sequence variants for the design of a high density 600 k SNP genotyping array for maize based on 57 M SNPs and small indels that were discovered by mapping whole genome sequencing reads of 30 representative temperate maize lines against B73 AGP_v2. For experimental validation, we selected 1.2 M variants by applying stringent filtering criteria. This 1.2 M subset was used to genotype 285 maize samples representing the genetic diversity of European (EU) and American (US) temperate maize as well as a sample of tropical maize lines. We created a final selection of 616,201 high quality variants based on their assay performance, physical distribution, and concordance with *in silico* variant calls from sequencing data. Here, we describe the design of the high density Affymetrix® Axiom® Maize Genotyping Array which represents a powerful tool for fine-mapping of genomic regions, genome-wide studies, and detection of marker-trait associations. We also demonstrate its application for investigating subpopulation structure and LD in diverse maize germplasm.

Results and discussion

Discovery and pre-selection of variants

For variant (i.e. SNP and indel) discovery we sequenced 30 maize inbreds composed of 17 European Flint lines as well as nine European and four US Dent lines (Additional file 1: Table S1). The lines represent important founder lines for maize breeding in Europe and the US and have been used in previous studies

[32,33]. Mapping the generated sequence reads to the B73 reference sequence (AGP_v2) resulted in 50-fold sequence coverage on average of four deep sequenced lines (DK105, EP1, F7, PH207) as well as 12-fold coverage on average of the 26 remaining lines. Based on the mapped sequence reads 56,938,462 variant positions were identified.

A filtered list of variants was created for quality score determination similar to the dual approach of Chia et al. [18]. Variants were included in this list if they were identified independently by two different programs, SAMtools [34] and GATK [35] and were characterized by high quality scores as well as presence of reference (B73) and non-reference alleles in the discovery panel. Applying these filters, the initial variant number was reduced by a factor of 10. We finally selected 5,593,169 bi-allelic variants for further analysis. 66.7% (3,731,960) of these variant positions were congruent with variants reported by [18] for the maize HapMap2 data. Of 46,660 variants from the Illumina® MaizeSNP50 BeadChip which could be uniquely anchored to the B73 reference sequence, 43,615 (93.5%) were also covered by *in silico* SNP calls from sequencing in our set of 5.6 M variants. This proportion is higher than the 72.3% overlap reported in the maize HapMap2 SNP dataset reported by [36] and can most likely be attributed to the higher sequence coverage in our study.

Selection of high-confidence variants for array construction

A multi-step filtering approach was applied to reduce the number of 5.6 M variants to a subset of 1.2 M variants for experimental validation on two Affymetrix® Axiom® 600 k screening arrays (Figure 1). From those, 616 k were selected for the design of the 600 k array.

Variant selection according to in-silico analysis of sequence data

The 5.6 M variants were filtered according to quality and their support by sequence reads. The sequenced lines were inbred lines with only minor residual heterozygosity (mean of 0.65%, Additional file 1: Table S2) as determined from Illumina® MaizeSNP50 data. In the 5.6 M variants, we observed 23.3% heterozygous compared to 72.7% homozygous calls, which was not expected from the Illumina® MaizeSNP50 genotyping data.

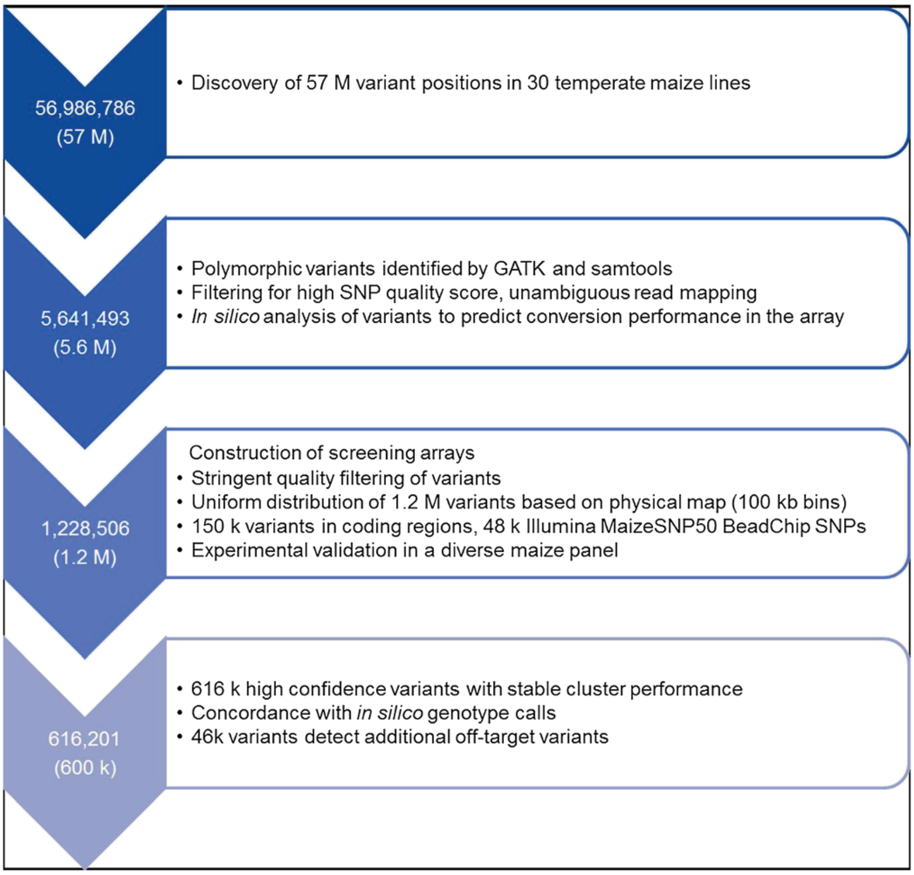
56,986,786
(57 M)
• Discovery of 57 M variant positions in 30 temperate maize lines
5,641,493
(5.6 M)
• Polymorphic variants identified by GATK and samtools
• Filtering for high SNP quality score, unambiguous read mapping
• *In silico* analysis of variants to predict conversion performance in the array
1,228,506
(1.2 M)
Construction of screening arrays
• Stringent quality filtering of variants
• Uniform distribution of 1.2 M variants based on physical map (100 kb bins)
• 150 k variants in coding regions, 48 k Illumina MaizeSNP50 BeadChip SNPs
• Experimental validation in a diverse maize panel
616,201
(600 k)
• 616 k high confidence variants with stable cluster performance
• Concordance with *in silico* genotype calls
• 46k variants detect additional off-target variants

Figure 1 Flow diagram with the major filtering steps. Flow diagram showing steps and major criteria of the variant selection process during development of the maize 600 k genotyping array.

Besides "true" heterozygous calls, such calls may arise from the large fraction of segmental duplications as well as orthologous and paralogous sequences retained in the ancient polyploid maize genome [15]. In line with this, the false discovery rate (FDR) of heterozygous calls was significantly higher (87.0%) compared to the FDR of homozygous calls (1.6%) as determined by comparison with variant calls from the Illumina® MaizeSNP50 BeadChip. Thus, in order to create a list of high quality variants only homozygous calls were considered for further analysis.

We decided to include all available 150,394 coding variants on the screening arrays, as these variants have a greater potential than non-coding variants to affect gene function. To enable comparison across studies, we further included 48,324 SNPs of the Illumina® MaizeSNP50 BeadChip as "must-have" variants. The remaining ~ 1 M positions on the screening arrays were filled with non-coding variants based on their distribution across the genome. Similar to the strategy reported by Kranis et al. [28], we applied a bin based approach with the intention to create a subset of physically equally distributed variants. We observed that variant numbers in centromeric bins were always lower than in telomeric bins, indicating lower polymorphism rates in the centromeric regions. This reduction of variant numbers around the centromeres was also observed in other maize studies [18,19,37] and may result from the high proportion of repetitive DNA around the centromeres for which no markers can be developed. Aiming simultaneously for a balanced representation of pool-specific as well as shared variants between Dent and Flint, 931,340 variants were included in the list for validation. We selected 158,448 additional variants to specifically increase the number of variants in under-represented bins to reach a final number of 1,228,506 variants which could be placed on the screening arrays. The marker density on the screening arrays was one variant per ~ 1.7 kb on average over all chromosomes (Additional file 1: Table S3).

Variant validation by genotyping 285 representative maize samples

In order to assemble a robust set of variants for design of the 600 k array, the selected set of 1.2 M variants was used to genotype 285 DNA samples from 280 diverse worldwide maize inbred lines and hybrids for the evaluation of variant performance (Additional file 1: Table S4). We investigated conversion performance of the variants on the array with respect to (i) genotype call rates, cluster separation, and reproducibility, (ii) polymorphism in the panel under study, and (iii) consistent Mendelian inheritance from parents to off-spring in trios.

Hybridization intensity signals were clustered by the Affymetrix Axiom GT1 algorithm and interpreted as homozygous, heterozygous, or no calls, respectively. Different from the situation in humans or animals, where samples are highly heterozygous, most of the samples in our maize validation panel were highly inbred. Thus, we compared genotype calls obtained with and without applying an inbred correction factor (Additional file 2: Figure S1). This factor was assigned to each sample to adjust the probability of observing a heterozygous call given the inbreeding level of the sample. The average call rate of the screening arrays could be increased by 2.3% to 98.1% upon inbred correction (Additional file 1: Table S5). With inbred correction, inbred line B73 exhibited the highest call rate (99.5%) and one F1 hybrid (UH007 x Lo11, 92.2%) together with Teosinte (acc. GID265285, 92.2%) the lowest call rates. Furthermore, American maize lines revealed higher call rates on average compared to European lines, followed by call rates of tropical lines and hybrids. This is in accordance with the literature [21] and suggests a negative correlation between call rate and increasing sequence divergence to the reference sequence of B73 from which probe sequences on the array were derived.

Based on genotype call cluster separation, cluster variance, and cluster position, variants were assigned to one out of six quality categories (Additional file 2: Figure S2). Comparing the category assignments with and without inbred correction resulted in a change of category in 36.2% of all variants (Additional file 1: Table S6). As expected, the category of variants fulfilling all cluster metric criteria and classified as "PolyHighResolution" (PHR) increased most, resulting in a gain of 30.7% upon inbred correction. Details on the number of variants from each category with and without inbred correction are given in Additional file 1: Table S6. In total, 25.1% of the newly developed 1,131,860 variants (excluding the Illumina® MaizeSNP50 variants) failed to convert and did not give reliable genotype calls upon inbred correction (designated "other" in Additional file 1: Table S6). The proportion of 74.9% converted variants is lower than in a similar study in chicken, where 82.0% of the variants could be converted into successful variants [28]. In rice which has an around five-fold smaller and less complex genome than maize, 84% of variants of the Illumina® RiceSNP50 array [8] were converted successfully (GenTrain score > 0.5). Given the higher complexity of the maize genome compared to chicken or rice, our conversion rate is in the expected range.

Selection of high-confidence variants and composition of the 600 k array

For the selection of high-confidence variants for the 600 k array, we applied a voting system based on (i) their performance on the screening arrays, (ii) concordance of array genotyping calls with *in silico* variant calls from sequencing data of the 30 maize lines in the discovery panel, and (iii) over- or under-representation of the corresponding bin. To ensure a high performance on the final array, the highest weight was assigned to the first criterion. We focussed on clearly separated genotype clusters with little variance that were not influenced by information regarding the inbreeding level (Additional file 2: Figure S1). Applying this procedure the 570,546 highest scoring variants as well as 45,655 SNPs of the Illumina® MaizeSNP50 BeadChip were included in the final selection for the 600 k array (Additional file 1: Table S6).

The 600 k genotyping array is composed of 616,201 variants (609,442 SNPs and 6,759 indels), corresponding to an average density of one variant per ~ 3.4 kb (median density one variant per 0.3 kb; Additional file 1: Table S3, Additional file 2: Figure S3). The average genetic distance between variants is 0.0025 cM, which corresponds to 406 variants per cM. The variants are evenly distributed across the chromosomes with the only exception of one region on the short arm of chromosome 6, where the maximal distance between neighboring variants exceeds 1.2 Mb. Despite a specific filter aiming for equal variant distribution according to the physical map distance, the final distribution followed the average recombination rate along chromosomes, which reflects varying polymorphism rates in the material under study (Figure 2). The highest density of variants was found in gene enriched telomeric regions, thus ensuring the maximal possible amount of genetic information in regions with high recombination rates. A comparable pattern of variant distribution as well as a lack of variants on the short arm of maize chromosome 6 in the nucleolus organizer region (NOR; approximate position 7–28 Mb) has been reported previously [18,19]. From the 616,201 variants represented on the Affymetrix® Axiom® Maize Array 561,751 (91.2%) are also present in the maize HapMap2 variants [18].

All 616,201 variant positions were annotated based on the B73 filtered gene set which comprises 39,656 genes (Additional file 1: Table S7), resulting in 26,620 genes (67.1%) tagged with at least one variant in their coding, intronic, or UTR region, compared to 17,520 genes tagged by SNPs of the Illumina® MaizeSNP50 BeadChip (44.2%). Including 5 kb up- and downstream regions, 35,089 genes (88.5%) were represented by at least one variant, thus providing an excellent basis for finding marker-trait associations in targeted and genome-wide approaches.

To determine the reproducibility of variants represented on the 600 k array, technical and biological replicates were analysed. First, three technical B37 replicates as internal controls exhibited up to 99.8% of identical genotype calls (Additional file 1: Table S8). Three biological replicates from different seed sources exhibited a high level of concordant genotype calls in the range of 99.76% to 99.84%. Furthermore, two lines (DK105 and EP1) were represented by two samples each comprised of a single plant and a pooled sample, respectively, showing 99.51% and 97.73% concordance. Some lack of concordance here can be explained by residual heterozygosity in the pooled samples. For determination of stable Mendelian inheritance, 23 trios with both parental lines as well as the corresponding F1 hybrid were analysed. These trios revealed stable Mendelian inheritance between parental lines and their offspring in 94.3% of the variants. After excluding the trio with the lowest call rate (UH007, Lo11, UH007 x Lo11) stable Mendelian inheritance could be observed in 97.6% of the variants, underlining the call rate as an indication of sample quality. The analysis of biological and technical replicates and trios confirmed the high reproducibility of genotype calls obtained with the variants represented on the Affymetrix® Axiom® Maize Array which is in the same range as reported for the Illumina® MaizeSNP50 BeadChip [21].

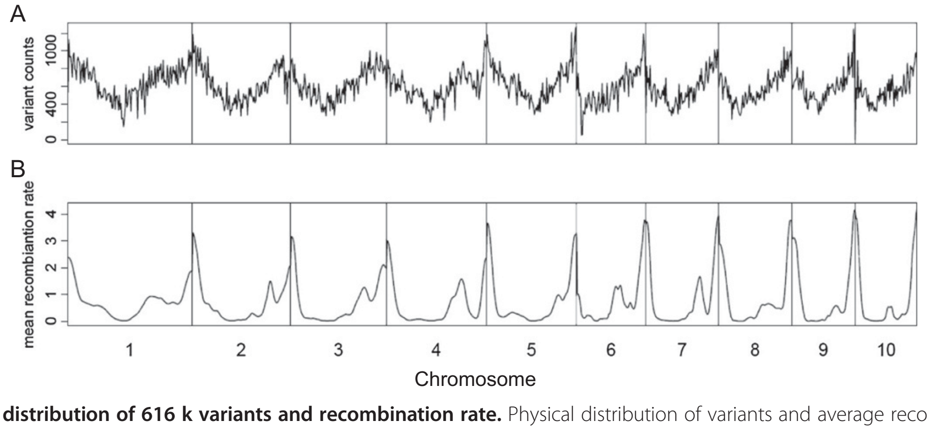

Figure 2 Physical distribution of 616 k variants and recombination rate. Physical distribution of variants and average recombination rate along the ten maize chromosomes depicted for 2 Mb windows. **A)** Distribution of 616 k variants represented on the 600 k array, **B)** Average recombination rate in cM/Mb from [32].

The usefulness of a genotyping array is characterized by the number of variants polymorphic in the panel of genotypes under study. In the 155 public maize lines, two Teosinte accessions, and 23 F1 hybrids used in this study for validation, 99.9% of the 600 k array variants were polymorphic. Only a small number of 262 variants (all derived from the Illumina® MaizeSNP50 BeadChip) were monomorphic across all samples of the validation panel. After excluding three genotypic samples without clear germplasm group assignment, 95.6% of the 600 k variants were polymorphic within Dent (N = 73), 98.7% in Flint (N = 79), and 97.2% within F1 hybrids (N = 23), respectively (Figure 3). Only 42.2% of the variants were polymorphic within the two Teosinte accessions. It must be noted however, that with only two samples the diversity in Teosinte is not well captured in our validation panel. Additionally, the array was not optimized for wild maize relatives as they were not included in the discovery panel. The high overall polymorphism rate depicts the quality of the filtering procedure and is in line or even exceeding results obtained by other studies regarding genotype array validation in animals and plants [8,21,38,39]. It confirms the utility of the array for a wide range of applications in maize germplasm.

Among the selected variants, one category called “Off-Target Variants” (OTVs) was of special interest since these 45,974 variants detect previously uncharacterized variants in the flanking region of the target variant. Due to a reduced hybridization efficiency OTVs are characterized by cluster splits or additional relatively low signal intensity clusters compared to expected homozygous and heterozygous genotypes (Additional file 2: Figure S2) and have been shown to be reproducible [40]. These 46 k variants offer the possibility not only to analyse the genotype call of the target variant, but provide in addition information on presence or absence of putative additional variants in the flanking regions. The latter information can be treated as a bi-allelic flanking variant and was included for population structure analyses.

Analysis of population substructure

The identification of population substructure is crucial for quantitative genetic or population genetic studies since population stratification or admixture may affect detection of marker-trait associations, genomic prediction, or estimation of population genetic parameters. To determine the ability of the variants represented on the maize 600 k genotyping array to resolve population structure, we performed principal coordinate (PCoA) and admixture analyses of 155 public inbred lines. The first principal coordinate revealed a clear separation of Dent and Flint with a small group of samples located in the center (Figure 4A). This central group included (sub)tropical Flint and Dent lines, the popcorn accession, two lines with unknown pedigree as well as Flint lines that originated from Southern Spain and one Flint line tracing back to Argentina. The clear separation of Dent and Flint reflects their genetic differentiation for more than 2,500 years [41], accompanied by varying adaptive and selective pressures. Similar results were obtained in studies based on isozymes [42], RFLPs [43,44], SSRs [41], and SNPs [45]. Analyzing the 73 Dent lines separately (Figure 4B), the first two axes further subdivided the samples into distinct subgroups, namely Iowa Stiff Stalk Synthetic (BSSS), Iodent, and Lancaster Sure Crop (LSC), with several non-BSSS and tropical lines in the center. The three groups BSSS, non-BSSS (including LSC), and Iodent represent three major heterotic groups

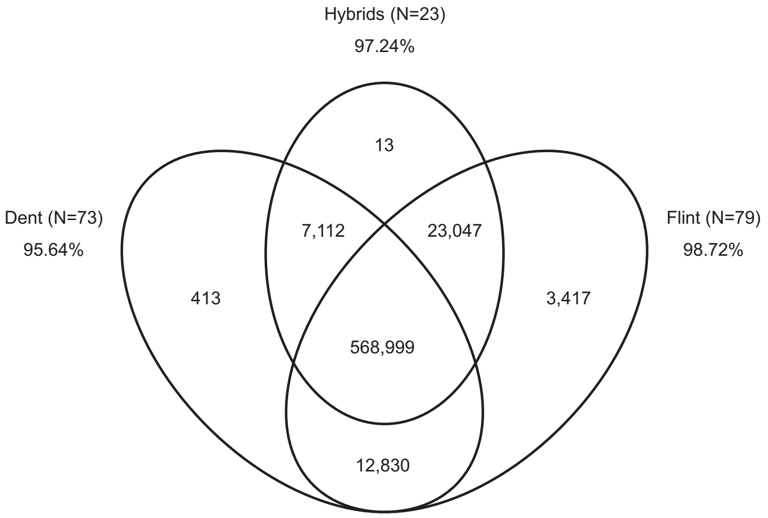

Figure 3 Polymorphic variants of the 600 k array. Venn diagram showing the number of polymorphic variants represented on the 600 k array in 73 Dent and 79 Flint samples and 23 hybrids of the validation panel.

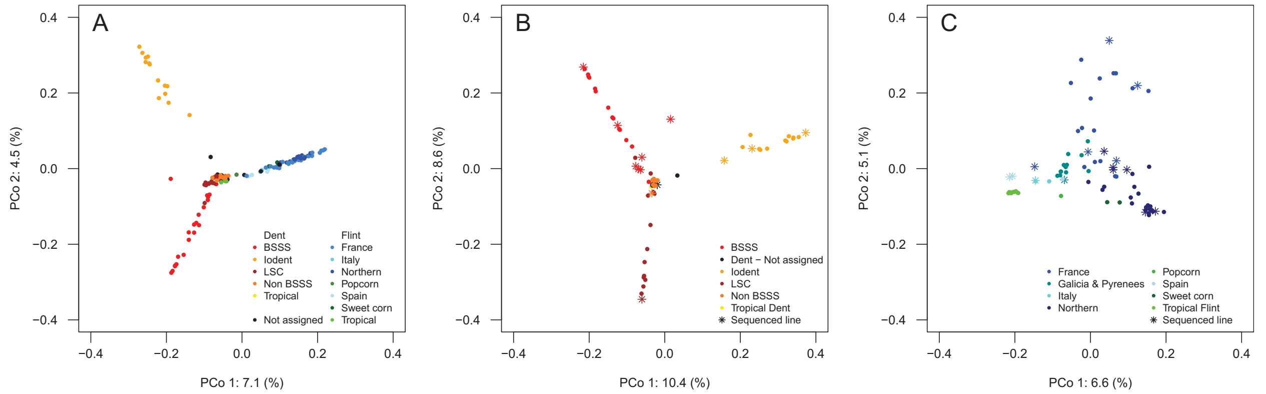

Figure 4 Population substructure in a diverse maize panel. PCoA plots of the first two axes in a diverse panel of public maize lines based on Rogers' distances from 251,152 variants including OTVs represented on the 600 k array (markers in LD with $r^2 > 0.8$ were excluded). **A)** Whole set of 155 maize lines, **B)** 73 Dent lines, and **C)** 79 Flint lines.

within temperate Dent, whose strong differentiation is well-known [46,47]. Compared to US material, European samples clustered within each group more towards the center (Figure 4B), suggesting a lower level of differentiation and population substructure [31].

In Central Europe, Flint plays an important role for hybrid breeding programs relying on the Dent x Flint heterotic pattern. PCoA of 76 Flint lines as well as one popcorn and two sweetcorn accessions resulted in the separation of European Flint lines adapted to more Northern or Mediterranean climate, respectively (Figure 4C). This split has also been observed in other studies based on phenotypic and RFLP marker data [48] and can be traced back to the introgression of maize to Europe. Maize was introduced to Europe starting at the end of the 15th century, when Columbus brought subtropical maize from the Caribbean Islands to Southern Spain, later followed by travelers importing so called Northern Flint [49] from Canada to Northern France [48,50-52]. The "non-Northern" Flint group in our study was further subdivided in the PCoA by the second axis depicting the relatedness of a subset of samples which had the French line F7 in their ancestry. Thus, the first two axes revealed two main subgroups of European Flint although the substructure was not as pronounced as in Dent. As indicated in Figure 4 (B, C), the sequenced lines of the variant discovery panel were nicely distributed across the different germplasm pools.

Cross-validation results obtained by ADMIXTURE [53] suggested K = 7 as the most likely number of groups (Additional file 2: Figure S4, Additional file 2: Figure S5) with four clear clusters in Dent for BSSS, Iodents, LSC, and a mixed group of non-BSSS lines, as well as two clusters for Northern Flints and non-Northern Flints, and a mixed group of (sub)tropical lines or lines with ancestors of (sub)tropical origin. This grouping well reflects the main subgroups observed with PCoA. In accordance with an increasing sequence divergence to the reference sequence of B73, ADMIXTURE analysis based on the 46 k flanking OTVs resulted in the subdivision of Dent, Flint, and a group including tropical lines as well as Flint lines originating from Argentina, Spain, and Italy (Additional file 2: Figure S6, Additional file 2: Figure S7).

We conclude that the variants represented on the 600 k array are well suited for dissecting the diversity and genetic composition of temperate maize lines. Performance of the array with regard to the analysis of tropical material or wild maize relatives will need further investigation.

Extent of linkage disequilibrium

The extent of LD in a population is influenced by recombination rate, drift, mutation, selection, and population structure. It has thus influence on experimental design, resolution, and analysis of genome-wide studies. In the public inbred lines genotyped for validation, LD decay ($r^2 \leq 0.2$) could be observed within 158 kb on average with some chromosomal differences (Table 1). Group specific analysis of the LD extent revealed a substantially higher level of LD in the two Dent groups of Iodents and BSSS with mean LD decay distances of 19.5 and 36.2 Mb, respectively, compared to non-BSSS lines (excluding the LSC group) where LD decayed within 239 kb. Due to the rather small sample size in LSC (N = 9), decay distances were not calculated for this subset. Mean LD decay values in Flint were highest for non-Northern Flints, which included several lines sharing a common ancestor, with 4.6 Mb, followed by Northern Flints (312 kb). The fastest LD decay was observed in (sub)tropical lines (70 kb). This corroborates previous reports supporting the close relationship and small number of founder lines within Iodent and BSSS compared to the other groups [19,47]. The low values of the non-BSSS as well as the

Table 1 Mean linkage disequilibrium (LD) given as r^2 and average LD decay distance[a] in kb per chromosome in 155 lines (all) and in six[b] subgroups as determined by ADMIXTURE

	All		Dent – BSSS		Dent – Non-BSSS		Iodent		Non-Northern Flint		Northern Flint		(Sub) Tropical	
	(N = 155)		(N = 14)		(N = 32)		(N = 14)		(N = 18)		(N = 34)		(N = 34)	
Chr.	mean r^2	r^2 decay	mean r^2	r^2 decay	mean r^2	r^2 decay	mean r^2	r^2 decay	mean r^2	r^2 decay	mean r^2	r^2 decay	mean r^2	r^2 decay
1	0.029	119.14	0.202	14,411.28	0.049	156.32	0.212	15,949.23	0.133	4,007.83	0.049	197.98	0.037	43.38
2	0.027	126.13	0.260	30,329.28	0.048	178.63	0.177	9,411.84	0.133	4,211.89	0.059	332.88	0.039	64.29
3	0.034	199.62	0.263	31,772.64	0.057	306.29	0.170	8,340.20	0.199	15,545.48	0.055	352.29	0.044	96.22
4	0.031	192.41	0.207	16,123.82	0.054	319.43	0.310	50,389.31	0.128	3,114.28	0.063	404.96	0.037	83.82
5	0.027	119.91	0.224	18,692.30	0.047	172.69	0.154	5,681.66	0.144	4,626.09	0.056	262.24	0.038	57.22
6	0.025	106.65	0.198	14,001.18	0.049	188.22	0.170	8,886.16	0.099	955.38	0.049	206.58	0.036	40.20
7	0.033	176.04	0.214	16,901.24	0.057	246.89	0.228	18,945.94	0.141	3,737.87	0.060	379.57	0.042	75.11
8	0.033	183.72	0.257	28,937.29	0.053	254.27	0.280	37,984.38	0.120	1,537.13	0.052	322.23	0.039	81.56
9	0.033	167.58	0.309	53,758.98	0.057	263.68	0.263	28,009.93	0.151	5,020.05	0.056	314.73	0.041	61.48
10	0.033	192.64	0.396	136,831.20	0.057	303.38	0.256	11,899.95	0.130	2,916.32	0.054	343.07	0.041	96.65
Mean	0.031	158.38	0.253	36,175.92	0.053	238.98	0.222	19,549.86	0.138	4,567.23	0.055	311.65	0.039	69.99
Median	0.032	171.81	0.241	23,814.80	0.054	250.58	0.220	13,924.59	0.133	3,872.85	0.055	327.55	0.039	69.70

[a] Distances in kb for $r^2 = 0.2$ calculated per 50 Mb window.
[b] LD decay distances were not calculated for LSC (N = 9).

(sub)tropical lines in our study might be explained by the high heterogeneity of both groups. Still, LD levels in our panel of maize lines were higher compared to previous studies reporting the breakdown of LD within distances between 5 and 10 kb [18,54] in highly diverse maize lines. The higher LD extent in our study might be due to the sample panel analysed, which mainly comprised temperate elite maize inbred lines belonging to distinct germplasm pools but no landraces or wild species. The variants selected for the 600 k array fulfill the requirements by [55], after which genotyping arrays should have sufficient coverage to capture the fastest LD decay of the considered heterotic pools. Thus, especially for analysis of diverse sample panels, high density genotyping arrays are of interest for estimating global and local LD.

Other potential applications of the maize high density array

We presented the usefulness of the maize high density 600 k array for the analysis of population structure and LD, but of course it is suitable for many other applications in maize research and breeding. For population genetic analyses based on genotyping data ascertainment bias is a central aspect [56]. The lines sequenced in this study were chosen to represent the diversity within the more comprehensive validation panel. However, bias may be introduced by the filtering steps that are applied during array development and typically results in a bias towards intermediate allele frequencies. Flanking OTVs have the advantage of not being directly targeted by the variant filtering procedure itself offering thus the potential to counteract ascertainment bias [40]. Compared to the minor allele frequency (MAF) distribution of the complete set of variants of the 600 k array where only 3.8% of the variants detect rare alleles with a MAF < 0.05, OTVs showed with 40.8% rare alleles (MAF < 0.05) a reduced bias towards intermediate allele frequencies (Figure 5). Thus, even if the specific type of the variant as well as its exact location is unknown, flanking OTVs

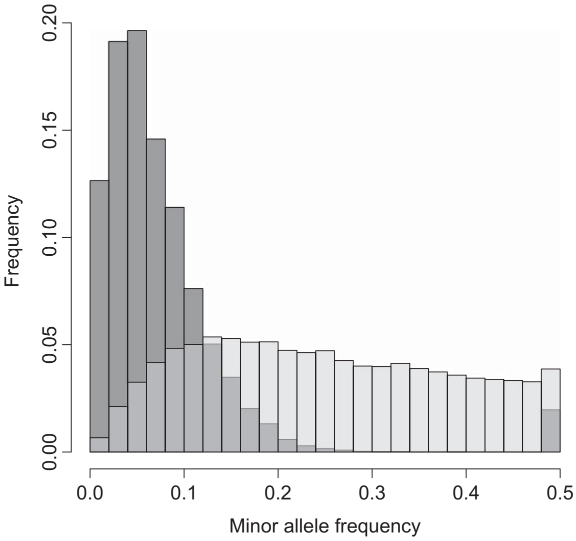

Figure 5 Minor allele frequency distribution for 616 k variants and 46 k flanking OTVs. Minor allele frequency (MAF) distribution in 155 maize lines for the 616 k variants (transparent grey) and for 46 k flanking OTVs (black).

represent an interesting subset of variants for population genetic analyses like screens for selection signatures based on genotyping data which we will address in further studies.

Further applications of the 600 k array include its use in genome-wide and targeted approaches. The array should have the desired density for genome-wide association studies in maize, for which the currently available density of the Illumina® MaizeSNP50 BeadChip was shown to provide limited resolution [57]. Due to the extremely high marker density, the array can be used for bulked segregant analysis to identify genomic regions involved in phenotypic traits with monogenic or oligogenic inheritance [58]. Further applications may be seen in the context of plant variety protection and in the investigation of pedigree relationships, identity-by-descent regions and ancestral lineages [59]. A panel of representative lines genotyped with the 600 k array should also allow high accuracy in imputation of genotypes from genetic material analysed with lower density marker panels [60]. Finally, custom sets of SNPs may be assembled from any genomic region and converted into other highly flexible SNP assay formats to saturate specific regions in fine-mapping, map-based cloning studies or marker-assisted selection, since flanking sequence information is available and conversion rates among SNP platforms are generally high.

Conclusions

This paper describes the establishment of the currently largest publicly available SNP array in crop species composed of 616,201 SNPs and small indels. The Affymetrix® Axiom® Maize Array is optimized for European and American temperate maize. It is well suited for fine-mapping of genomic regions, genome-wide studies, and detection of marker-trait associations. Important aspects in the development of the maize 600 k genotyping array were: (i) identification of polymorphic, high-confidence variants based on whole genome sequence data of 30 representative temperate maize lines, (ii) selection of physically equally distributed variants for validation, taking predicted variant performance and subgroup specific segregation into account, (iii) experimental variant validation by genotyping 285 DNA samples originating from diverse subgroups of maize, and (iv) final selection of variants upon stringent filtering based on cluster metrics, concordance with *in silico* calls, and physical position. We have shown that the variants selected for the 600 k genotyping array were polymorphic in a broad maize panel, ensuring its suitability for a wide range of applications. We investigated a subset of variants (OTVs) that showed almost no bias towards intermediate allele frequencies, thus are potentially of interest for population genetic analyses. Finally, we performed principal coordinate, admixture, and LD analyses to illustrate the potential of the array to analyse population substructure and LD decay with high resolution.

Methods

Sequencing of 30 maize lines in the variant discovery panel

For whole genome sequencing of 30 maize inbred lines (Additional file 1: Table S1), DNA was extracted from leaf material frozen in liquid nitrogen following the protocol of [61]. For deep-sequencing of the three Flint lines DK105, EP1, and F7, as well as the Dent line PH207, DNA was extracted from a single plant each. For sequencing of the other 26 maize lines, DNA was extracted from bulked leaf samples of 8–10 plants per line. Sequencing was performed on an Illumina® HiSeq 2000 platform, generating 2x100 bp paired-end reads from standard 300 bp libraries using manufacturer's protocols.

In silico variant discovery and pre-selection

Sequence reads were mapped against reference sequence B73 AGP_v2 using BWA (version 0.7.5) [62]. Alignments were post-processed by marking duplicates and fixing paired-end information applying the PICARD toolbox (version 1.84) [63] and by performing local realignments using the Genome Analysis Toolkit (GATK, version 2.4-9) [35]. Quality scores of called variant positions (SNPs and short insertions/deletions) were improved by recalibrating values according to results of several covariate analyses (homopolymer, cycle, dinucleotide, quality score) done on a set of trusted variants. As no SNP database was available for the maize varieties under study, a database of high quality trusted SNPs was created following the recommendation on the GATK website (http://www.broadinstitute.org/gatk/guide/tagged?tag=baserecalibrator).

Briefly, initial variants were called independently using two algorithms to obtain a more robust SNP set, as recommended by [64]. We used SAMtools (version 0.1.18) [34] and the intersection to the initial GATK variants was further filtered for SNP quality (≥ 50); low and excessive read coverage ($50 \leq DP < 3000$); presence of the reference allele; and homozygous non-reference calls in at least two of the 30 lines. In a second round, variants were identified from the base quality recalibrated bam files by the GATK Unified Genotyper. For the final set of candidate variants, several stringent filters were applied. First, variants were excluded if they were located in regions with genomic copy number ≥ 50 (based on 16-kmer counts). Second, variants were not forwarded to the next step if (i) more than 5% of reads had mapping quality 0, (ii) coverage was more than six fold higher compared to the mean coverage, or (iii) a SNP quality score was below 100. In addition, variants had to exhibit a minimal distance of 20 bp between neighboring variants located in at least one flanking sequence. In summary, the final pre-selection variant set scored for tiling by the Affymetrix® Axiom® myDesign GW bioinformatics pipeline comprised a total of 5,641,493 bi-allelic variants. For annotation of the variants, version 5b60 of the reference sequence B73 AGP_v2 was used (ftp://ftp.gramene.org/pub/gramene/maizesequence.org/release-5b/filtered-set/) which contains 39,656 gene models. Variant effects were predicted using SNPeff (version 3.2) [65].

Variant selection for the screening arrays according to predicted conversion quality, physical position, and segregation in Dent and Flint

For all variants from the 5.6 M list p-convert values were calculated per probe according to the Affymetrix® Axiom myDesign GW bioinformatics pipeline and categorized as "recommended", "neutral", "not recommended", and "not possible", respectively. The p-convert value can take a value between 0 and 1 and describes the predicted probability to convert on the array by taking its sequence, binding energies, expected degree of non-specific binding and hybridization to multiple genomic regions into account. Two probe sets (forward and reverse) for each SNP from the Illumina® MaizeSNP50 BeadChip (GenTrain score > 0) were directly included in the list of variants for the screening arrays without further filtering unless they were classified as "not possible". For the newly identified variants only probe sets categorized as "recommended" or "neutral" were further analyzed. For coding variants, the probe with the higher p-convert value was chosen based on this classification, whereas for all remaining variants probe sets were further filtered according to the following multi step approach. Based on the reference genome size of 2.066 Gb, first, the maize genome was partitioned in 20,660 bins of size 100 kb, aiming at an equal physical distribution of variants. Assuming up to 1.23 M possible variants, which could be tested on two screening arrays, after substracting the fixed variants (150 k coding variants and 2*48 k Illumina® MaizeSNP50 BeadChip SNPs), each 100 kb bin would contain on average 48 variants. Three cases were distinguished to fill the physical bins: (i) all possible variants of a bin were included if less than 48 "recommended" or "neutral" variants were identified in the corresponding bin, (ii) "recommended" variants were considered as fixed, if their number did not exceed 48 and remaining "neutral" variants were subjected to another filtering step, and (iii) "recommended" variants were further filtered, if ≥ 48 were observed in the corresponding bin. For this filtering step to fill up underrepresented bins, allele frequencies were determined for Dent (N = 13) and Flint (N = 17) lines separately by calculating the ratio of homozygous non-reference allele calls in relation to all available calls per variant. Variants were classified according to their pool-specific allele frequency as class "A", corresponding to intermediate (between 0.2 and 0.8), or "B" to extreme non-reference allele frequencies (< 0.2 or > 0.8). One third of the variants, which filled up the bins, were chosen to be specific for Flint (category Dent "A" | Flint "B") and one third specific for Dent (Dent "B" | Flint "A"), respectively. Further, one sixth each had to be either common (category Dent "A" | Flint "A") or rare for both groups (Dent "B" | Flint "B"), respectively.

In a final step, 50 kb bins were considered if the original 100 kb bin had been filled with less than 48 variants. Additional variants were selected if there were less than 8 variants per 50 kb bin to avoid underrepresentation of genomic regions by choosing variants randomly (i) from "recommended" variants with at least six, but less than 22 homozygous reference allele calls in at least 28 of the 30 lines of the discovery panel to avoid extreme allele frequencies (maximal six variants per 50 kb bin), and (ii) if further variants were required, from all remaining variants. Altogether, a final list of 1,228,506 variants was established for validation with a diverse panel of maize lines on two customized 675 k Affymetrix® Axiom® myDesign GW screening arrays.

Plant material for genotyping

The selected set of 1.2 M variants was used to genotype 285 DNA samples from genetically diverse maize germplasm to evaluate their assay performance. The validation panel was composed of 224 Dent and Flint inbred lines of which 92 were proprietary lines. From those, line B37 was included three times as technical replicate and three lines (B73, DK105, EP1) were represented by two biological replicates each. In addition, 13 tropical lines (ten Flint, three Dent), ten doubled haploid lines from three European Flint landraces, four lines with no available pool assignment, and two Teosinte accessions were analysed. Finally, we included 27 hybrids, among which there were 23 F1 hybrids from Mendelian trios with both parental lines present in the public elite line panel, and four proprietary hybrids (Additional file 1: Table S4). The Dent elite lines comprised representative samples belonging to the subgroups Iowa Stiff Stalk Synthetic (BSSS), Lancaster Sure Crop (LSC), or Iodent, as well as other non-BSSS samples, and samples with tropical origin. The Flint panel was composed of European Northern Flints and lines originating from Spain, Italy, and France, as well as sweetcorn, popcorn, and tropical lines. Except for the 92 proprietary inbred lines, the elite inbred lines were selected according to their frequency of use and citation [46,47] as well as based on utilization in other studies, pedigree information or classifcation available from literature [66,67] or from internet sources [68] with the aim to represent diverse temperate material. The 96 proprietary samples were included in the analysis of the screening array for training of the variant clustering algorithm, but not in further analyses presented here.

DNA for genotyping was extracted from seeds available to the authors or kindly provided by the following institutions: INRA UMR de Génétique Végétale (Gif-sur-Yvette, France), Universität Hohenheim (Stuttgart, Germany), USDA-ARS (Ames, USA), CIAM (La Coruña, Spain), CRA-MAC Maize Research Unit (Bergamo, Italy), and CSIC (Pontevedra, Spain).

Comparison of variant calls with the Illumina® MaizeSNP50 BeadChip

The 30 sequenced lines (Additional file 1: Table S1) were genotyped with the Illumina® MaizeSNP50 BeadChip following manufacturer's protocols using a total of 50 ng genomic DNA. Raw hybridization intensity data processing, clustering, and genotype calling were performed using the software GenomeStudio (v2011.1, Illumina®) and the public cluster file II described in [21].

Experimental variant validation by genotyping

From each sample, 200 ng genomic DNA per array was used for analysis on the Affymetrix GeneTitan® platform with the Axiom myDesign GW genotyping array following manufacturer's protocol. After array processing, four samples were excluded from further analyses as signal intensity files could be created for only one of the two screening arrays, resulting in 281 samples remaining for further investigation (Additional file 1: Table S4).

Raw hybridization intensity data processing, clustering, genotype calling (genotypes AA, AB, BB), off-target variant (OTV; genotypes AA, AB, BB, OO) calling, and variant categorization according to genotype cluster metrics (Additional file 2: Figure S2) were performed using Affymetrix Power Tools (APT, version 1.15.0) and the package SNPolisher (version 1.3.6.6) [69] for R (version 3.0.1) [70] according to the Axiom Genotyping Solution Data Analysis Guide. For initial genotype calling generic *a priori* cluster positions were used since no information about expected cluster positions was available. The three possible genotype clusters were then redefined in *a posteriori* cluster positions, taking the observed genotype call positions into account and variants were finally classified according to selected cluster metrics. A first analysis was performed according to the recommendations of Affymetrix, but with a reduced threshold (0.90) for the variant call frequency instead of the default value (0.97) to account for the high amount of PAVs in the maize genome [17].

In a second, extended analysis different levels of inbreeding were taken into account for *a posteriori* cluster definition because of the high amount of lines in the validation panel exhibiting only a small proportion of heterozygosity in contrast to populations in Hardy-Weinberg equilibrium. The inbred correction was achieved by a parameter assigning sample-specific penalties using the "–read-inbred" parameter for the "apt-probeset-genotype" command in APT. This parameter takes values from 0 for fully heterozygous to 16 for completely homozygous samples and includes this information for re-defining *a priori* cluster positions for genotype calling. We assigned values of 0 for F1 hybrids, 12 for inbreds with unclear homozygosity level, and 14 for pure inbred and doubled haploid lines to allow some remaining heterozygosity (Additional file 1: Table S4). Results of the analyses with and without inbred correction were compared and a subset of randomly selected genotype clusters were visually checked.

Selection of high-confidence variants for construction of the final 600 k array

Variants were preferentially selected if they were exhibiting stable category assignments (Additional file 2: Figure S2) with clearly separated clusters to avoid restrictions dependent on the inbred-level. Categories were assigned by the classification step of SNPolisher using the following parameters: CR.cut = 90, FLD.cut = 3.6, HetSO.cut = −0.1, HetSO.OTV.cut = −0.3, HomRO2.cut = 0.3, HomRO3.cut = −0.9, HomRO.flag = TRUE, nMinorAllele.cut = 2. For high quality variant selection, a total of 523,154 variants classified as "PolyHighResolution" (PHR) with and without inbred correction were directly forwarded to the final list as they were characterized by distinct and narrow clusters in both analyses. These variants were used to define customized cluster quality criteria for OTVs to ensure a clear separation of genotype clusters, but allowing in addition lower heterozygous cluster signal intensities due to cluster splits caused by unexpected off-target variants in the flanking region of the target variant or potential tri-allelic variants. The "Fisher´s Linear Discriminant" (FLD) value characterized the cluster quality being highest in case of well-separated and narrow clusters. The "Heterozygous Cluster Strength Offset" (HetSO) measured the difference in the signal intensities of the genotype clusters as the heterozygous cluster should have higher signal intensity on average compared to the homozygous ones due to technical features of the array. The "Homozygote Ratio Offset" (HomRO) described the distance of the homozygous clusters to the heterozygous one to detect potentially misplaced clusters. The chosen thresholds upon inbred correction were the following: no monomorphic variants, ≤ 10% missing calls (corresponding to ≤ 30 missing calls), FLD > 3.5, HetSO > −3.5, and HomRO > 1. As FLD and HetSO values were exhibiting missing values in some variants with only two clusters, an additional threshold was set in this case using a FLD value between homozygous clusters (homFLD) of > 5. All 42,877 variants which were classified in both analyses (with and without inbred correction) as OTV and passed in addition the above thresholds were included in the selection for the final array.

Remaining variants were ranked by applying a voting system. First, variants were ranked according to their classification with and without inbred correction. Variants, which were classified as "OTVstable" or changed their category from "NoMinorHom" (only one homozygous and one heterozygous genotype cluster) without inbred correction to PHR after inbred correction, were assigned a weight of 10. Variants, which belonged to any other class without inbred correction, but changed to PHR after inbred correction received a weight of 5, and all remaining variants not fulfilling the previous criteria obtained a weight of 0. Second, variants were weighted regarding the concordance of their calls with the *in silico* variant calls from sequencing. The number of matching calls per variant across the 30 sequenced lines from the discovery panel, which all were also analyzed on the genotyping test arrays, was normalized to the total number of calls per variant resulting in a value in the range of 0 to 1. As a third criterion, the over- or underrepresentation of the corresponding 100 kb bin was taken into account by calculating the deviation of the number of variants in the corresponding bin to the mean of variants in the five bins up- and downstream, respectively, and scaling the value into a range of values between −1 and 1. For the final rank, the sum was built of (i) the weight of the variant class that was multiplied with 35 to ensure a high performance on the final array (range: 0 to 750), (ii) the value of the sequence match multiplied with 90 to minimize false-positives (range: 0 to 90), and (iii) the weight of the bin representation, which was multiplied by 10 for lowest impact (range: −10 to 10). For the 48,324 Illumina® MaizeSNP50 SNPs which were tiled from both sides, the probe with the higher rank was included in the final set in case of varying ranks. If both probes of a variant exhibited the same rank, one probe was chosen randomly. Due to an erroneous mapping of 2,669 Illumina® MaizeSNP50 BeadChip SNPs to the B73 reference sequence a wrong (non-polymorphic) position was assayed on the screening arrays and these non-validated SNPs were not included on the final array. The top 616,201 variants were selected for the final array design among which 45,655 originated from the Illumina® MaizeSNP50 BeadChip. Information on SNP IDs, genome positions, probe sets, and alleles are available at NCBI GEO as platform GPL18778 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL18778) or from the product information of manufacturer Affymetrix.

Analyses of population substructure and linkage disequilibrium

For all analyses indels were treated as bi-allelic SNPs. In PCoA and ADMIXTURE analyses OTVs were included with their genotype calls of the target variant as well as information on presence or absence of a flanking variant, resulting in 616 k plus 46 k variants. Variants with ≥ 10% of missing data were excluded. Remaining missing data were imputed using Beagle [71] via the R package "synbreed" [72] with R version 3.0.1 [70]. Public inbreds (N = 155, replicates excluded) of the validation panel were investigated with PCoA and ADMIXTURE for population structure as well as for LD decay. LD pruning was performed for PCoA and ADMIXTURE analyses by applying a r^2 threshold of 0.8. PCoA based on Rogers´ distances was performed using R with the packages "synbreed" [72], "adegenet" [73], and "ape" [74]. Analysis of population substructure was calculated using ADMIXTURE (version 1.23) [53] running with default settings for K = 1 to K = 15. LD was calculated chromosome-wise per 50 Mb window using Plink (version 1.07) [75] and LD decay analysis was performed using the R package "synbreed" [72].

Availability of supporting data

Supporting sequence data are available in the NCBI Sequence Read Archive (SRA) repository under BioProject accession number PRJNA260788 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA260788). Information on SNP IDs, genome positions, probe sets, and alleles can be retrieved from NCBI GEO, platform GPL18778 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL18778).

Additional files

Additional file 1: Table S1. Description of the sequence variant discovery panel with group assignment, origin (Europe or USA), raw sequence coverage, number of replicates per line, and assigned inbred penalty (cf. genotype calling in Material and Methods section). **Table S2.** Percentage of heterozygous and missing calls, respectively, for samples of the discovery panel calculated from 49,574 Illumina® MaizeSNP50 BeadChip SNPs with a GenTrain score > 0. **Table S3.** Number of variants as well as median, mean, and maximal distance between neighboring variants in kb per chromosome and mean genetic distance in cM for screening arrays and final array, respectively. **Table S4.** Description of validation panel with group assignment, origin (CA: Canada, EU: Europe, MX: Mexico, SA: South Africa, US: United States of America, "-": no information available), source of the material (proprietary: plant material from KWS SAAT AG), inbred penalty, and number of replicates. **Table S5.** Call rates of validation samples (N = 281) on the two screening arrays, without and with inbred correction. Samples are sorted according to mean call rate with inbred correction across arrays (last column). **Table S6.** Number and percentage of variants per category with and without inbred correction for new identified and Illumina® MaizeSNP50 BeadChip variants, respectively, on the screening arrays and on the 600 k array. **Table S7.** Annotation and prediction of variant effects for the 616,201 variants of the maize 600 k array. Predictions were obtained with SNPeff [65]. Multiple entries per variant are possible. **Table S8.** Overview of replicates included in the validation panel and corresponding percentage of genotype call concordance calculcated from 570 k SNPs (omitting variants with flanking OTVs).

Additional file 2: Figure S1. Effects of inbred correction on genotype calling in predominantly homozygous inbred lines shown for two variants. **Figure S2.** Representative cluster plots for the six categories according to SNPolisher. **Figure S3.** Variant density shown for the screening arrays (light grey) and for the variants of the Affymetrix® Axiom® Maize Array (black) across the 10 maize chromosomes. Centromere positions are indicated by a black horizontal bar. **Figure S4.** Cross-validation errors from ADMIXTURE for different values of K for 155 maize lines based on 251,152 variants including OTVs (markers in LD with $r^2 > 0.8$ were excluded). **Figure S5.** Subgroups identified in 155 maize lines of the validation panel as revealed by ADMIXTURE for K = 7 based on 251,152 variants including OTVs (markers in LD with $r^2 > 0.8$ were excluded). **Figure S6.** Cross-validation errors from ADMIXTURE for different values of K for 155 maize lines based on 27,099 flanking OTVs (markers in LD with $r^2 > 0.8$ were excluded). **Figure S7.** Subgroups identified in the 155 public lines of the validation panel as revealed by admixture for K = 3 based on 27,099 flanking OTVs (markers in LD with $r^2 > 0.8$ were excluded).

Competing interests

Two commercial organisations – KWS SAAT AG and Affymetrix Inc. – were involved in the development of the array and in preparation of the manuscript. Among the authors, CK and MO work for KWS SAAT AG, and AD and CB work for Affymetrix.

Authors' contributions

This study was carried out in collaboration between all authors. CCS, EB, GH, KFXM, and MO conceived the study; CK and EB provided DNA for sequencing and genotyping as well as information on the maize lines; TM and TMS sequenced the DNA samples of the discovery panel; GH and MS carried out variant calling and bioinformatic analyses; CB performed the Affymetrix® Axiom® design; CCS, CK, EB, GH, MS, and SU developed and discussed the variant selection strategy and discussed results; HP and RF processed the genotyping arrays; AD helped with the genotype calling pipeline; EB, GH, MS, and SU analyzed the genotypic data; EB and SU drafted the manuscript; CCS, GH, and MS edited the manuscript and all authors read and approved the final manuscript.

Acknowledgements

We are grateful to Alain Charcosset (INRA UMR de Génétique Végétale, Gif-sur-Yvette, France), Albrecht E. Melchinger (Universität Hohenheim, Institute of Plant Breeding, Seed Science and Population Genetics, Stuttgart, Germany), Mark J. Millard (USDA-ARS, Ames, USA), Jesús Moreno González (CIAM, La Coruña, Spain), Rita Redaelli (CRA-MAC Maize Research Unit, Bergamo, Italy), Pedro Revilla (CSIC, Pontevedra, Spain), and Karl Schmid (Universität Hohenheim, Crop Plant Biodiversity and Breeding Informatics) for providing seeds of maize lines and information about pedigrees and group assignment. This study was funded by the Federal Ministry of Education and Research (BMBF, Germany) within the AgroClustEr Synbreed - Synergistic plant and animal breeding (grant 0315528) and publication was supported by the German Research Foundation (DFG) and the Technische Universität München within the funding programme Open Access Publishing.

Author details

[1]Plant Breeding, Centre of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising, Germany. [2]Plant Genome and System Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [3]KWS SAAT AG, 37555 Einbeck, Germany. [4]Institute of Human Genetics, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [5]Animal Breeding, Centre of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising, Germany. [6]Affymetrix Inc., Santa Clara, CA 95051, USA.

Received: 11 June 2014 Accepted: 23 September 2014
Published: 29 September 2014

References

1. Ragoussis J: **Genotyping technologies for genetic research.** *Annu Rev Genomics Hum Genet* 2009, **10:**117–133.
2. Hayes BJ, Lewin HA, Goddard ME: **The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation.** *Trends Genet* 2013, **29**(4):206–214.
3. Langridge P, Fleury D: **Making the most of 'omics' for crop breeding.** *Trends Biotechnol* 2011, **29**(1):33–40.
4. Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR: **Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*.** *Nat Commun* 2011, **2:**467.
5. Myocardial Infarction Genetics Consortium: **Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants.** *Nat Genet* 2009, **41**(3):334–341.
6. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, 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 C, Todesco M: **Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines.** *Nature* 2010, **465**(7298):627–631.
7. Hackett CA, McLean K, Bryan GJ: **Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population.** *PLoS ONE* 2013, **8**(5):e63939.
8. Chen H, Xie W, He H, Yu H, Chen W, Li J, Yu R, Yao Y, Zhang W, He Y, Tang X, Zhou F, Deng XW, Zhang Q: **A high-density SNP genotyping array for rice biology and molecular breeding.** *Mol Plant* 2014, **7**(3):541–553.
9. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**(4):1819–1829.
10. Rincent R, Laloe D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodriguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schön CC, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L: **Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.).** *Genetics* 2012, **192**(2):715–728.
11. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H: **A genome-wide scan for signatures of recent selection in Holstein cattle.** *Anim Genet* 2010, **41**(4):377–389.
12. McEvoy BP, Montgomery GW, McRae AF, Ripatti S, Perola M, Spector TD, Cherkas L, Ahmadi KR, Boomsma D, Willemsen G, Hottenga JJ, Pedersen NL, Magnusson PKE, Kyvik KO, Christensen K, Kaprio J, Heikkilä K, Palotie A, Widen E, Muilu J, Syvänen A-C, Liljedahl U, Hardiman O, Cronin S, Peltonen L, Martin NG, Visscher PM: **Geographical structure and differential natural selection among North European populations.** *Genome Res* 2009, **19**(5):804–814.
13. van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, Gonzalez JDS, Ross-Ibarra J: **Genetic signals of origin, spread, and introgression in a large sample of maize landraces.** *Proc Natl Acad Sci U S A* 2011, **108**(3):1088–1092.
14. Ahn S, Tanksley SD: **Comparative linkage maps of the rice and maize genomes.** *Proc Natl Acad Sci U S A* 1993, **90**(17):7980–7984.
15. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326**(5956):1112–1115.
16. Schnable JC, Springer NM, Freeling M: **Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss.** *Proc Natl Acad Sci U S A* 2011, **108**(10):4069–4074.
17. Springer NM, Ying K, Fu Y, Ji TM, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS: **Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content.** *PLoS Genet* 2009, **5**(11):e1000734.
18. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, Gore M, Guill KE, Holland J, Hufford MB, Lai J, Li M, Liu X, Lu Y, McCombie R, Nelson R, Poland J, Prasanna BM, Pyhajarvi T, Rong T, Sekhon RS, Sun Q, Tenaillon MI, Tian F, Wang J, Xu X, *et al*: **Maize HapMap2 identifies extant variation from a genome in flux.** *Nat Genet* 2012, **44**(7):803–807.
19. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA: **Comprehensive genotyping of the USA national maize inbred seed bank.** *Genome Biol* 2013, **14**(6):R55.
20. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.** *PLoS ONE* 2011, **6**(5):e19379.
21. Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent

P, Rose M, Schön CC, Sun Q, Walter H, Martin OC, Falque M: **A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome.** *PLoS ONE* 2011, **6**(12):e28334.

22. Li H, Peng ZY, Yang XH, Wang WD, Fu JJ, Wang JH, Han YJ, Chai YC, Guo TT, Yang N, Liu J, Warburton ML, Cheng YB, Hao XM, Zhang P, Zhao JY, Liu YJ, Wang GY, Li JS, Yan JB: **Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels.** *Nat Genet* 2013, **45**(1):43–50.
23. Gresset S, Westermeier P, Rademacher S, Ouzunova M, Presterl T, Westhoff P, Schön C-C: **Stable carbon isotope discrimination is under genetic control in the C4 species maize with several genomic regions influencing trait expression.** *Plant Physiol* 2014, **164**(1):131–143.
24. Albrecht T, Auinger HJ, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho HP, Schön C-C: **Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years.** *Theor Appl Genet* 2014, **127**(6):1375–1386.
25. Hufford MB, Lubinksy P, Pyhajarvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J: **The genomic signature of crop-wild introgression in maize.** *PLoS Genet* 2013, **9**(5):e1003477.
26. Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, Cregan PB: **Development and evaluation of SoySNP50K, a high-density genotyping array for soybean.** *PLoS ONE* 2013, **8**(1):e54985.
27. Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieseke R, Plieske J, International Wheat Genome Sequencing C, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, *et al*: **Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array.** *Plant Biotechnol J* 2014, 12(6):787–796.
28. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, Talbot R, Pirani A, Brew F, Kaiser P, Hocking P, Fife M, Salmon N, Fulton J, Strom T, Haberer G, Weigend S, Preisinger R, Gholami M, Qanbari S, Simianer H, Watson K, Woolliams J, Burt D: **Development of a high density 600K SNP genotyping array for chicken.** *BMC Genomics* 2013, **14**:59.
29. Rincon G, Weber KL, Eenennaam AL, Golden BL, Medrano JF: **Performance of bovine high-density genotyping platforms in Holsteins and Jerseys.** *J Dairy Sci* 2011, **94**(12):6116–6121.
30. Matukumalli LK, Schroeder S, DeNise SK, Sonstegard TS, Lawley CT, Georges M, Coppieters W, Gietzen K, Medrano JF, Rincon G, Lince D, Eggen A, Glaser L, Cam G, Van Tassel C: **Analyzing LD blocks and CNV segments in cattle: novel genomic features identified using the BovineHD BeadChip.** In Pub No 370-2011-002. San Diego, CA: Illumina Inc; 2011.
31. van Heerwaarden J, Hufford MB, Ross-Ibarra J: **Historical genomics of North American maize.** *Proc Natl Acad Sci U S A* 2012, **109**(31):12420–12425.
32. Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincent R, Schipprack W, Altmann T, Flament P, Melchinger AE, Menz M, Moreno-Gonzalez J, Ouzunova M, Revilla P, Charcosset A, Martin OC, Schön CC: **Intraspecific variation of recombination rate in maize.** *Genome Biol* 2013, **14**(9):R103.
33. Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, Flament P, Melchinger AE, Menz M, Meyer N, Moreau L, Moreno-González J, Ouzunova M, Pausch H, Ranc N, Schipprack W, Schönleben M, Walter H, Charcosset A, Schön C-C: **Usefulness of multi-parental populations of maize (*Zea mays* L.) for genome-based prediction.** *Genetics* 2014, **198**(1):3–16.
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20**(9):1297–1303.
36. Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner RA, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J: **Comparative population genomics of maize domestication and improvement.** *Nat Genet* 2012, **44**(7):808–811.
37. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES: **A first-generation haplotype map of maize.** *Science* 2009, **326**(5956):1115–1117.
38. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadieu-Tonon C, Eggen A, Heuven HCM, Jamli S, Jiken AJ, Klopp C, Lawley CT, McEwan J, Martin P, Moreno CR, Mulsant P, Nabihoudine I, Pailhoux E, Palhiere I, Rupp R, Sarry J, Sayre BL, Tircazes A, Wang J, Wang W, Zhang WG, Consortium IGG: **Design and characterization of a 52K SNP chip for goats.** *PLoS ONE* 2014, **9**(1):e86227.
39. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu Z-L, Kerstens HH, Law AS, Megens H-J, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PLoS ONE* 2009, **4**(8):e6524.
40. Didion JP, Yang HN, Sheppard K, Fu CP, McMillan L, de Villena FPM, Churchill GA: **Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias.** *BMC Genomics* 2012, **13**:34.
41. Labate JA, Lamkey KR, Mitchell SE, Kresovich S, Sullivan H, Smith JSC: **Molecular and historical aspects of corn belt dent diversity.** *Crop Sci* 2003, **43**(1):80–91.
42. Doebley J, Wendel JD, Smith JSC, Stuber CW, Goodman MM: **The origin of Cornbelt maize - the isozyme evidence.** *Econ Bot* 1988, **42**(1):120–131.
43. Messmer MM, Melchinger AE, Boppenmaier J, Brunklaus-Jung E, Herrmann RG: **Relationships among early European maize inbreds. 1. Genetic diversity among flint and dent lines revealed by RFLPs.** *Crop Sci* 1992, **32**(6):1301–1309.
44. Dubreuil P, Dufour P, Krejci E, Causse M, de Vienne D, Gallais A, Charcosset A: **Organization of RFLP diversity among inbred lines of maize representing the most significant heterotic groups.** *Crop Sci* 1996, **36**(3):790–799.
45. Frascaroli E, Schrag TA, Melchinger AE: **Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs.** *Theor Appl Genet* 2013, **126**(1):133–141.
46. Mikel MA: **Availability and analysis of proprietary dent corn inbred lines with expired US plant variety protection.** *Crop Sci* 2006, **46**(6):2555–2560.
47. Mikel MA, Dudley JW: **Evolution of North American dent corn from public to proprietary germplasm.** *Crop Sci* 2006, **46**(3):1193–1205.
48. Rebourg C, Gouesnard B, Charcosset A: **Large scale molecular analysis of traditional European maize populations. Relationships with morphological variation.** *Heredity* 2001, **86**:574–587.
49. Brown WL, Anderson E: **The Northern flint corns.** *Ann Missouri Bot Garden* 1947, **34**:1–28.
50. Dubreuil P, Warburton M, Chastanet M, Hoisington D, Charcosset A: **More on the introduction of temperate maize into Europe: Large-scale bulk SSR genotyping and new historical elements.** *Maydica* 2006, **51**(2):281–291.
51. Rebourg C, Chastanet M, Gouesnard B, Welcker C, Dubreuil P, Charcosset A: **Maize introduction into Europe: the history reviewed in the light of molecular data.** *Theor Appl Genet* 2003, **106**(5):895–903.
52. Tenaillon MI, Charcosset A: **A European perspective on maize history.** *C R Biol* 2011, **334**(3):221–228.
53. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19**(9):1655–1664.
54. Yan JB, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J: **Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers.** *PLoS ONE* 2009, **4**(12):e8451.
55. Van Inghelandt D, Reif JC, Dhillon BS, Flament P, Melchinger AE: **Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm.** *Theor Appl Genet* 2011, **123**(1):11–20.
56. Nielsen R: **Population genetic analysis of ascertained SNP data.** *Human Genomics* 2004, **1**(3):218–224.
57. Zila CT, Samayoa LF, Santiago R, Butrón A, Holland JB: **A genome-wide association study reveals genes associated with *Fusarium* ear rot resistance in a maize core diversity panel.** *G3: Genes Genomes Genet* 2013, **3**(11):2095–2104.
58. Becker A, Chao D-Y, Zhang X, Salt DE, Baxter I: **Bulk segregant analysis using single nucleotide polymorphism microarrays.** *PLoS ONE* 2011, **6**(1):e15993.
59. Thompson EA: **Identity by descent: variation in meiosis, across genomes, and in populations.** *Genetics* 2013, **194**(2):301–326.
60. Pausch H, Aigner B, Emmerling R, Edel C, Gotz K-U, Fries R: **Imputation of high-density genotypes in the Fleckvieh cattle population.** *Genet Sel Evol* 2013, **45**(1):3.

61. Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW: **Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics.** *Proc Natl Acad Sci U S A* 1984, **81**(24):8014–8018.
62. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754–1760.
63. **PICARD, A set of tools (in Java) for working with next generation sequencing data in the BAM format.** [http://broadinstitute.github.io/picard/]
64. Yu X, Sun S: **Comparing a few SNP calling algorithms using low-coverage sequencing data.** *BMC Bioinformatics* 2013, **14**:274.
65. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3.** *Fly* 2012, **6**(2):80–92.
66. Nelson PT, Coles ND, Holland JB, Bubeck DM, Smith S, Goodman MM: **Molecular characterization of maize inbreds with expired US plant variety protection.** *Crop Sci* 2008, **48**(5):1673–1685.
67. Flint-Garcia SA, Thuillet AC, Yu JM, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Maize association population: a high-resolution platform for quantitative trait locus dissection.** *Plant J* 2005, **44**(6):1054–1064.
68. **USDA ARS National Genetic Resources Program. Germplasm Resources Information Network - (GRIN) [Online Database].** [http://www.ars-grin.gov/cgi-bin/npgs/html/crop.pl?89]
69. Pirani A, Gao H, Bellon L, Webster T: **Best practices for genotyping analysis of plant and animal genomes with Affymetrix® Axiom® arrays.** In *The International Plant & Animal Genome XXI Conference.* San Diego, CA, USA: Scherago International; 2013:P0997. SNPolisher™ is an R package available from Affymetrix Inc. and can be downloaded from the "DevNet Tools" on www.affymetrix.com.
70. R Core Team: **R: A language and environment for statistical computing.** 2013, http://www.R-project.org/.
71. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *Am J Hum Genet* 2009, **84**(2):210–223.
72. Wimmer V, Albrecht T, Auinger HJ, Schön C-C: **synbreed: a framework for the analysis of genomic prediction data using R.** *Bioinformatics* 2012, **28**:2086–2087.
73. Jombart T: **adegenet: a R package for the multivariate analysis of genetic markers.** *Bioinformatics* 2008, **24**(11):1403–1405.
74. Paradis E, Claude J, Strimmer K: **APE: analyses of phylogenetics and evolution in R language.** *Bioinformatics* 2004, **20**:289–290.
75. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: A tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**(3):559–575.

doi:10.1186/1471-2164-15-823
Cite this article as: Unterseer *et al.*: A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 2014 **15**:823.

Unterseer *et al. Genome Biology* (2016) 17:137
DOI 10.1186/s13059-016-1009-x

Genome Biology

RESEARCH Open Access

A comprehensive study of the genomic differentiation between temperate Dent and Flint maize

Sandra Unterseer[1], Saurabh D. Pophaly[2], Regina Peis[1], Peter Westermeier[1,3], Manfred Mayer[1], Michael A. Seidel[4], Georg Haberer[4], Klaus F. X. Mayer[4], Bernardo Ordas[5], Hubert Pausch[6], Aurélien Tellier[2], Eva Bauer[1] and Chris-Carolin Schön[1]*

Abstract

Background: Dent and Flint represent two major germplasm pools exploited in maize breeding. Several traits differentiate the two pools, like cold tolerance, early vigor, and flowering time. A comparative investigation of their genomic architecture relevant for quantitative trait expression has not been reported so far. Understanding the genomic differences between germplasm pools may contribute to a better understanding of the complementarity in heterotic patterns exploited in hybrid breeding and of mechanisms involved in adaptation to different environments.

Results: We perform whole-genome screens for signatures of selection specific to temperate Dent and Flint maize by comparing high-density genotyping data of 70 American and European Dent and 66 European Flint inbred lines. We find 2.2 % and 1.4 % of the genes are under selective pressure, respectively, and identify candidate genes associated with agronomic traits known to differ between the two pools. Taking flowering time as an example for the differentiation between Dent and Flint, we investigate candidate genes involved in the flowering network by phenotypic analyses in a Dent–Flint introgression library and find that the Flint haplotypes of the candidates promote earlier flowering. Within the flowering network, the majority of Flint candidates are associated with endogenous pathways in contrast to Dent candidate genes, which are mainly involved in response to environmental factors like light and photoperiod. The diversity patterns of the candidates in a unique panel of more than 900 individuals from 38 European landraces indicate a major contribution of landraces from France, Germany, and Spain to the candidate gene diversity of the Flint elite lines.

Conclusions: In this study, we report the investigation of pool-specific differences between temperate Dent and Flint on a genome-wide scale. The identified candidate genes represent a promising source for the functional investigation of pool-specific haplotypes in different genetic backgrounds and for the evaluation of their potential for future crop improvement like the adaptation to specific environments.

Keywords: Maize, Flint, Dent, Selection, Population genetics, Genomics, Genome-wide screen, Landraces

* Correspondence: chris.schoen@tum.de
[1]Plant Breeding, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany
Full list of author information is available at the end of the article

Background

Maize is one of the world's major staple crops but considerable concern is arising that ongoing anthropogenic global warming will have drastic effects on maize production and might result in a reduction of up to 10 % in yield in the near future [1]. Expanding production areas to higher latitudes could moderate the effect, but this would require the adaptation of breeding material to shorter vegetation periods. Breeders can cope with this challenge by taking advantage of the tremendous genetic diversity of maize that is available in different temperate breeding pools. Two of the major pools exploited in breeding are the Dent and Flint germplasm pools with their names referring to different kernel phenotypes [2]. Dents have characteristic indented kernels with high soft starch content, whereas Flints have kernels with a thick, hard, and vitreous outer layer (Fig. 1a). The genetic divergence of these two pools can be explained by their historic geographical separation [3] and adaptation to different environments. Among all maize germplasm, Northern Flints reached the highest latitudes like the northern regions of the U.S. and Canada, which required selection for early maturity and cold tolerance [3]. These Northern Flints, together with Caribbean germplasm, were major progenitors of European maize and enabled the rapid adaptation to European climates [4]. Especially in cooler regions of Europe, breeding programs exploit heterotic effects between Dent lines tracing back to U.S. Corn Belt Dents and Flint lines, with Flint contributing early vigor and good cold tolerance and Dent contributing

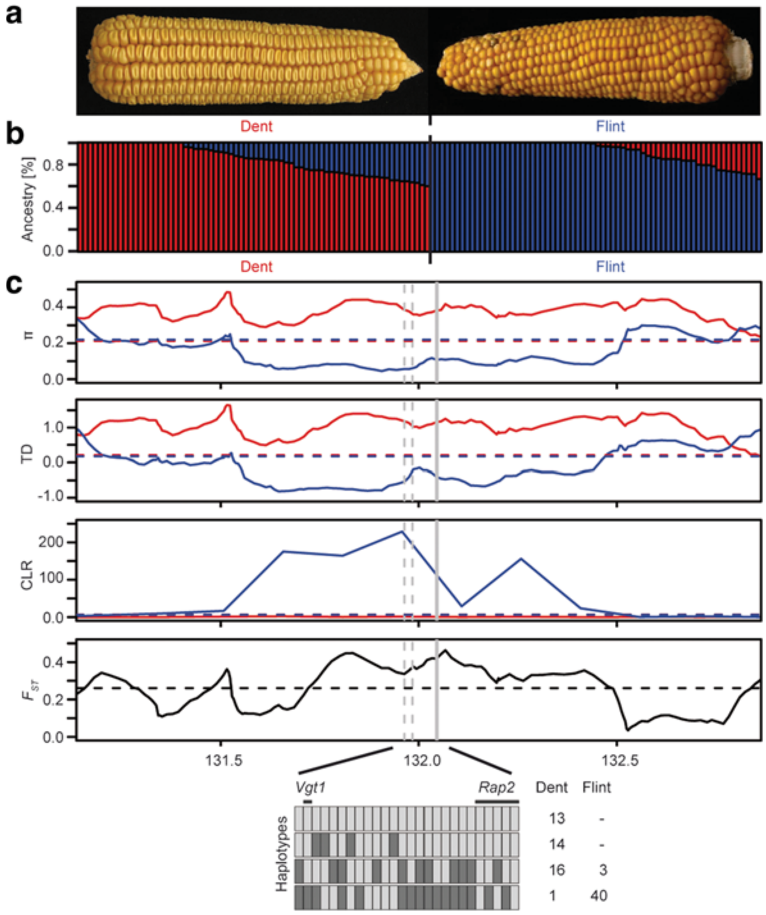

Fig. 1 Population structure of the investigated 136 Dent and Flint elite lines and detection of pool-specific selection signatures. **a** *Images* of maize cobs with Dent-type (*left*) and Flint-type kernels (*right*) as an example for phenotypic differences between the two germplasm pools. **b** Population structure and assignment of 136 temperate maize elite lines to Dent (*red*; N = 70) and Flint (*blue*; N = 66) pools. *Bar plots* indicate the relative ancestral composition of the lines. **c** Sweep statistics based on the panel of 136 temperate inbred lines shown exemplarily for a region on chromosome 8 that includes the *Vgt1* locus (*dashed gray lines*) and *Rap2* (*solid gray line*). Within-group statistics (*π*, TD, and CLR) are shown in *red* for Dent and in *blue* for Flint. *Horizontal dashed lines* indicate the cutoff per statistic (10 % quantile for *π* and TD, 90 % quantile for CLR and F_{ST}). For the region encompassed by the two loci *Vgt1* and *Rap2*, the four major haplotypes observed in the panel are shown. *Light gray boxes* indicate the B73 reference allele and *dark gray boxes* the alternative allele of each SNP. *Numbers on the right side* of the haplotype plot refer to the number of observations per haplotype within the Dent and the Flint panels

high productivity to the hybrids. The divergence of the Dent and Flint germplasm pools has been described in diversity studies based on molecular markers [5] and also in genetic studies mapping quantitative trait loci (QTL) underlying agronomic traits. A recent study utilizing Dent and Flint nested association mapping (NAM) populations [6] found little overlap of QTL for five complex traits between the two pools [7]. Although QTL mapping is a useful tool to elucidate the genetic architecture of phenotypic traits, it can only unravel genomic regions for which the genetic material under study is segregating, whereas regions under selection can be missed in case of near or complete fixation. Thus, alternative approaches are needed to investigate the divergence of Dent and Flint on a genomic level and to further elucidate how selection shaped the pool-specific genomic diversity.

Selection creates specific patterns of diversity in the genome [8] and these signatures can be used for the detection of regions under selection. When a favorable, new (derived compared to the ancestral) allele rises in frequency within a population, selective sweeps are generated, which are characterized by a local reduction in nucleotide diversity and high derived allele frequencies [9–11]. In addition, strong and recent sweeps will display large blocks with high linkage disequilibrium surrounding the derived mutation as the dispersal of the new allele will be faster than recombination is able to break down linkage disequilibrium [12, 13]. The identification of selection signatures through genome-wide screens provides an efficient way to detect selection candidates and methods for their detection are often combined to reduce the number of false-positives [14–16]. In maize, genome-wide screens for selection signatures were successfully applied to identify genes involved in domestication and improvement and allowed insights into evolutionary processes shaping the genome diversity of maize [17–20]. Taking advantage of the characteristics of selective sweeps and using high-density genotyping data from a maize 600 k single nucleotide polymorphism (SNP) array [21], we screened a panel of 136 temperate Dent and Flint elite lines for extreme allele frequencies over extended linked sites to identify genomic regions under selective pressure and to gain insights into the genomic variation underlying the differentiation of Dent and Flint. We included outgroup information from *Sorghum bicolor* to further support the identified candidate genes based on derived allele frequencies. We furthermore investigated the candidate genes based on whole-genome sequence data of 40 Dent and Flint lines [21, 22] and examined if genic and upstream regions contributed equally to the differentiation between temperate Dent and Flint.

The elite line panel under study comprised frequently used and important founder lines exploited in breeding programs for temperate climates. The Dent lines in our panel represent U.S. Corn Belt and European material, whereas most of the Flint lines originated from European breeding programs. Based on the selection screens, we examined pool-specific enrichment of candidate genes for metabolic pathways and investigated candidates associated with traits that are known to differentiate Dent and Flint like cold tolerance and flowering time [23, 24]. Flowering time is essential for local adaptation and represents a major determinant for other agronomic traits, such as grain filling and yield. The complex genetic architecture of flowering time has been studied in maize in a large number of studies mapping QTL with a meta-QTL analysis revealing 62 flowering time consensus QTL [25]. Phenotypic differences in maize flowering time are mainly caused by the accumulation of many small-effect QTL [26] and only a few large-effect genes have been characterized so far [27–30]. Hundreds of homologs to *A. thaliana* flowering time genes have been found in the maize genome [31], but in most cases their functional roles in the maize flowering network remain to be elucidated [25, 26, 29, 32–37]. In this study, we identified candidate genes from the flowering network with haplotypes near fixation or fixed in either of the two elite pools. We used this set of genes as an example to characterize genomic differentiation between Dent and Flint in more detail. We evaluated the effect of these genes on flowering time in a Dent–Flint introgression library and investigated their assignment to different pathways within the flowering network. To assess the congruency of the allelic composition of the candidate genes between elite lines and landraces, we expanded our candidate gene analysis to a large dataset of 38 European landraces that comprises more than 900 individuals. By exploring this unique resource, we gained insights into the genetic variation of the selection candidates between landraces and elite lines and investigated, which landraces likely contributed to the observed candidate gene diversity in the elite lines and if haplotypes not yet exploited in breeding could be detected. Taken together, our study allowed insights into patterns of differentiation between temperate Dent and Flint germplasm and provided candidates for follow-up studies to characterize their biological and molecular functions, to investigate their impact on phenotypes, and to assess their potential use for further crop improvement.

Results and discussion

Characterization of the Dent and Flint panels

We genotyped a diverse panel of 136 temperate inbred lines (Additional file 1: Table S1) at high density with the Axiom® Maize Genotyping Array [21]. The array comprises more than 600 k SNP markers, which were identified based on mid- to high-coverage whole-genome sequence data of 30 representative temperate Dent and

Flint maize lines [21]. Markers were filtered according to quality scores and stable performance on the array, thus representing high-confidence sequence variants, and their final distribution followed the average recombination rate along the chromosomes [21]. After stringent quality filtering of the 616,201 markers included on the array, 547,412 high-quality SNPs (88.8 %) remained for analysis. These SNPs tagged 19,759 genes (49.8 % of the annotated gene set of maize) with, on average, two SNPs in their coding region (52.6 % synonymous and 47.4 % non-synonymous). Slightly more SNPs were polymorphic in the Flint compared to the Dent panel (95.4 % versus 93.1 %), but the majority of SNPs segregated in both germplasm pools (88.6 %).

The panel of 136 temperate Dent and Flint inbred lines comprised frequently used and important founder lines exploited in breeding programs in Europe and the U.S., including lines which were used as parents for the U.S. and European NAM panels [6, 38, 39]. The 70 Dent lines were selected according to available pedigree information and their frequency of use and citation [40, 41] to assemble a representative set of lines. Besides 16 European Dent lines, the lines represent U.S. Corn Belt Dent and include lines from the Maize Association Population [42] and the list of inbred lines with expired U.S. plant variety protection [43]. The 66 Flint lines investigated in this study comprised important founder lines of European breeding programs like F2 and F7 originating from the French landrace Lacaune, EP1 from the Spanish landrace Lizargarate, and derivatives of the German landrace Gelber Badischer Landmais [44]. The Flints comprised in total 34 lines from France, 20 from Germany, four from Spain, three from Italy, three from North America, as well as one from Switzerland and Austria. Between the elite lines of the two germplasm pools, we observed a clear separation of pools (Fig. 1b) and a high genome-wide level of differentiation ($F_{ST} = 0.14$), which is consistent with the long-term genetic differentiation between Dent-type and Flint-type maize [2, 3].

Genome-wide screens for selection signals

Taking advantage of the characteristics of selective sweeps, we screened the genome for extreme allele frequencies over extended linked sites to detect regions under differential selective pressure between Dent and Flint. Signatures of selection in only one of the two pools, Dent or Flint, were detected based on low levels of nucleotide diversity (π) [9] and Tajima's D (TD) [10] in the respective pool. In addition, a signature had to be supported by a high value of the composite likelihood ratio (CLR) test [11] within the respective pool, which indicates a deviation of the allelic composition of a genetic region compared to a neutrally evolving sequence determined by the genomic background. To ensure that the selection signature was specific for one of the two pools, it had to be associated with a high level of differentiation between Dent and Flint measured by the fixation index F_{ST} [45]. Except for the CLR statistic, which was calculated for non-overlapping grids of 150 kb, we applied a sliding window approach averaging data over windows of 40 SNPs (sliding by 10 %) and filtered for regions below the 10 % quantile for π and TD and above the 90 % quantile for F_{ST} and CLR (Additional file 1: Table S2). Following the approach reported by [17], adjacent windows passing the threshold for all four statistics were grouped together for candidate gene analysis, as the observed changes in allele frequency were likely caused by the same selective sweep event. This resulted in a filtered set of 265 windows for Dent and 158 windows for Flint, with an average length of 331.40 kb and 267.80 kb, respectively, and thus comparable to the length of domestication windows found in a previous study [17]. An example of a signature of differential selection in Dent and Flint determined by all four metrics (π, TD, CLR, and F_{ST}) is shown in Fig. 1c for a region on chromosome 8 harboring two candidate genes. The underlying genetic region was composed of four major haplotypes. The first three haplotypes occurred at intermediate frequencies in Dent, whereas the fourth haplotype was almost exclusive for Flint.

Genome-wide patterns of diversity and the resulting distribution of selection signatures in the Dent and Flint panels are given in Additional file 2: Figure S1. Within the filtered set of windows, which covered 4.3 % of the total length of the maize genome for Dent and 2.1 % for Flint, we identified 876 genes as candidates under differential selective pressure in Dent and 545 genes for Flint with 14 genes common to both candidate genes sets (Additional file 3: Table S3). This corresponded to 2.2 % and 1.4 % of the filtered gene set of maize, respectively, and is in the same order of magnitude as the estimated number of genes under selective pressure during maize domestication and improvement [17]. When comparing the candidate gene sets with the 571 improvement candidates reported by [17], 26 genes overlapped with the list of Dent candidates but only one gene with the Flint candidate gene set. Considering that the genetic material studied in [17] comprised mainly U.S. Dent and (sub-) tropical lines and that pool-specific sequence variation in temperate Dent and Flint has been reported here and, for example, by [5], these results emphasize the relevance of a representative panel of lines belonging to divergent germplasm pools to obtain a comprehensive picture of the genomic diversity in maize.

In genome-wide screens for signatures of positive selection, also other forces than selection, such as heterogeneous mutation and recombination rates along the genome, past demographic history and background selection shape the genomic diversity and can give rise to false-positive signals. It is beyond the scope of this paper to infer a full demographic history of maize for the elite lines and landraces as the breeding history of maize is complex and violates several assumptions of the classic population genetics models (e.g. discussed in [46]), as, for example, the assumption of panmictic populations and applicability of the coalescent at short time scales. We therefore applied the CLR test [11], which detects selective sweeps based on the comparison of the site-frequency spectrum within a specific genomic region to the average site-frequency spectrum over the genome, a method which has been successfully used in human and other species to detect selective sweeps [11, 47, 48]. To further decrease the rate of false-positives, the CLR test was combined with three additional metrics (π, TD, and F_{ST}) and we identified signatures of positive selection based on this conservative approach with an overlap of genome-wide extreme values per metric. The high level of linkage disequilibrium in temperate Dent and Flint elite lines [21] facilitates the detection of selective sweep signals over sufficiently large genomic regions by the CLR test. On the other hand, the extent of linkage disequilibrium may decrease the power to discriminate between signals caused by genetic hitchhiking due to positive selection and negative background selection in regions with reduced levels of recombination [49, 50]. To assess the number of false-positives due to this effect, we explored the recombination landscape in the Dent and Flint panels by estimating lower bounds of historical recombination events [51]. The proportion of candidate genes located in regions with strongly reduced recombination rates and high linkage disequilibrium like (peri-) centromeric regions was then estimated. We found that 74.8 % of the Dent and 80.9 % of the Flint candidates were not located in regions with low levels of recombination (10 % quantile per chromosome; Additional file 1: Figure S2) indicating that the majority of candidates represent targets of selection rather than false-positive signals. Furthermore, in a classic selective sweep scenario (in contrast to background selection) targets of selection are to be enriched for derived alleles. As an additional test of our candidate regions, we included information from *Sorghum bicolor* to distinguish between ancestral and derived alleles. The Dent and Flint candidate gene sets revealed significantly higher derived allele frequencies compared to the remaining genes as measured by Fay and Wu's normalized H [52] ($p < 2.2\text{e-}16$; Additional file 1: Table S4), which also supported positive selection as the driving force of the observed allele frequency changes.

Gene ontology and pathway analyses of candidate gene sets

Considering genetic differentiation and distinct phenotypic characteristics of Dent and Flint, we tested whether the candidate gene sets were enriched for specific biological processes or pathways. Gene ontology (GO) terms associated with the identified genes were available for around 40 % of the candidates (333 for Dent and 214 for Flint). No significant GO term enrichment of biological processes, cellular components, and molecular functions could be detected for either of the two sets (Additional file 1: Figure S3). To investigate if candidate genes revealed a pool-specific enrichment for metabolic pathways, we performed pathway analyses using MapMan [53]. Based on information available for 58 Dent and 40 Flint candidate genes, we observed a grouping of genes associated with tetrapyrroles (chlorophyll and heme precursors) for Dent and for terpenoid metabolism for Flint (Additional file 1: Figure S4). The latter included the two genes *ZmPPS7.3* (*GRMZM2G014508*) and *ZmPPS8.2* (*GRMZM2G483889*), which encode a large and a small subunit of the geranyl diphosphate synthase complex in maize, respectively [54]. Like their homologues in *A. thaliana* [55], they are assumed to be involved in the biosynthesis of precursors of hormones from the isoprenoid pathway (e.g. gibberellins, brassinosteroids, and abscisic acid). The ability to produce other downstream products of this enzyme, namely β-caryophyllenes, has been shown to differ between European Flint and U.S. Dent lines and suggested that this defense response signal against herbivores was largely lost in temperate U.S. Dent [56, 57]. The analysis of candidates associated with other traits that are known to differentiate Dent and Flint revealed six Flint candidates that, according to GO terms, are related to cold tolerance, a trait that is characteristic for temperate Flint [24]. For two of the candidates, differential expression upon exposure to chilling temperature has been reported in maize (*GRMZM2G035584* [58] and *GRMZM2G095562* [59]) as well as for the homologous gene of *GRMZM2G139680* in rice [60]. The molecular and functional characterization of the identified candidate genes in maize and the investigation of differences between Dent and Flint in the regulation of phytohormone pathways or secondary metabolism may provide further insights in the adaptation of maize to different environments. Up to now, comprehensive RNA expression data across various developmental stages and tissues are mainly available for U.S. Dent lines like B73, which underlines the need for a better structural and functional genomic characterization of the Flint germplasm pool and its unique properties.

Assessing the phenotypic effects of candidate genes on flowering time in a Dent–Flint introgression library

In the genome-wide selection screens, we identified 18 candidates for Dent and 12 candidates for Flint, which could be assigned to the flowering pathway based on previous reports in maize, GO terms, and/or sequence homology to flowering genes characterized in other species [30, 32–34, 61, 62]. We focused exemplarily on candidate genes associated with the flowering network in maize as flowering time is an important agronomic trait that differentiates temperate Dent and Flint. However, functional studies of these genes in maize were available for only 30 % of the candidates (Additional file 4: Table S5). Here, we investigated the effect of the flowering time candidate genes in more detail using a maize introgression library.

The introgression library had a Dent genetic background with introgressions from a Flint donor line and comprised 97 lines, which carried single Flint segments and covered in total 50.9 % of the Flint donor genome (1048.7 Mb) with a median length of the donor genome segment size of 10.6 Mb (average: 30.8 Mb; Additional file 5: Table S6). We obtained phenotypic data for male and female flowering time based on a field experiment carried out at two locations in Germany. Heritabilities were 0.60 ($CI_{0.95}$ = [0.40; 0.73]) and 0.51 ($CI_{0.95}$ = [0.27; 0.67]) for male and female flowering time, respectively. Phenotypic differences between the Dent and Flint parent were larger for male than for female flowering time (23.2 and 17.8 days, respectively). Based on the least significant difference ($\alpha = 0.05$), 63 (64.9 %; Fig. 2a) and 16 lines (16.5 %; Fig. 2b) differed significantly from the recurrent Dent parent for male and female flowering time, respectively. Fifteen of these lines had significant effects for both male and female flowering time ($\alpha = 0.05$). When correcting for multiple testing ($\alpha = 0.05/97$), six lines (6.2 %) differed significantly for male and none for female flowering time.

Of the 97 lines, 22 carried a Flint introgression harboring one or several of the flowering time candidates identified in the selection screens (Additional file 5: Table S6). Fourteen of the 30 candidates were represented in these 22 introgression lines. Seven lines carried a segment with one or more of seven flowering time candidates identified in Flint and nine lines carried one or more of six flowering time candidates identified in Dent. Six lines carried a segment with a combination of Dent and Flint candidates. Although 75 lines did not carry one of the flowering time candidates identified in our selection screens, they may carry other flowering time genes with alleles differing between the Dent and the Flint parent of the introgression library. Lines carrying the Flint haplotype of a Flint candidate differed significantly from the 75 lines which did not carry one of the flowering time candidates from the selection screens (93.1 versus 96.1 days, p value = 0.011; Fig. 2a). For the lines which carried the Flint haplotype of a Dent selection candidate, this difference was not significant. The results indicate that in the genetic material under study, the Flint haplotypes of Flint candidates promoted flowering time more than the Flint haplotypes of Dent candidates.

Of the six lines with significant difference in male flowering compared to the Dent parent after correcting for multiple testing ($\alpha = 0.05/97$), two carried Flint haplotypes of Flint candidates and one a Flint haplotype of a Dent candidate. One of the lines included the well-characterized large-effect region comprising the ethylene-responsive transcription factor *Rap2* (related to *APETALA2 7*, *ZmRap2.7*, *Rap2*, *GRMZM2G700665*) and its regulatory upstream locus *Vgt1* [62], a major QTL for flowering time in maize [29, 30]. The other line contained *Zcn1* (one of several members of the *ZEA CENTRORADIALIS* or *TERMINAL FLOWER1* (*TFL1)*-like gene family [32]; also *Phosphatidylethanolamine-binding protein1*, *Pebp1*, *GRMZM2G092008*), for which so far only a moderate effect on flowering time was reported in maize [63]. This gene is related to *TFL1* in *A. thaliana* [32], which is an

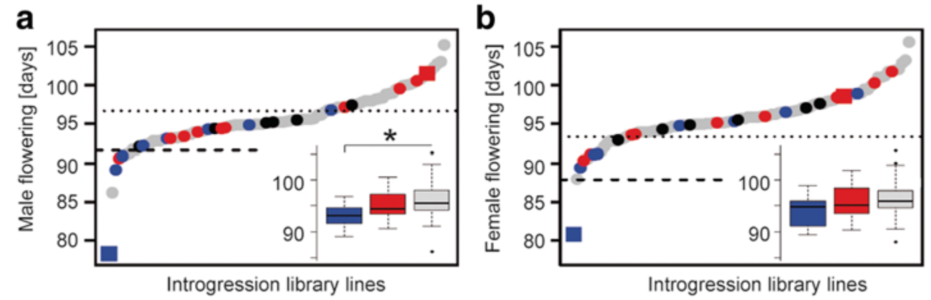

Fig. 2 Effect of candidate genes on flowering time in a Dent–Flint introgression library. Adjusted means of (**a**) male and (**b**) female flowering times for 97 introgression lines (*circles*) and the Dent and the Flint parental line (*red and blue squares*, respectively). Lines carrying a segment with Dent or Flint flowering time candidate genes are highlighted in *red* or *blue*, respectively, and lines with a Dent and a Flint candidate are shown in *black*. The *dotted and dashed lines* represent the significance thresholds without ($\alpha = 0.05$) and with correction for multiple testing ($\alpha = 0.05/97$). *Boxplots* of adjusted means of flowering times are depicted in the lower parts of (**a**) and (**b**) for seven lines carrying Flint haplotypes of Flint flowering time candidates (*blue*), nine lines carrying Flint haplotypes of Dent flowering time candidates (*red*), and the 75 lines not carrying a flowering time candidate (*gray*). For details about the respective lines see Additional file 5: Table S6. *Boxplots* show the upper and lower quartile, median (*horizontal bar*), and whiskers (*vertical bars*) of the adjusted means. Points above and below the whiskers indicate values ± 1.5 times the interquartile range. Significance of Student's *t*-tests with $p < 0.05$ is indicated by *

antagonist of the *FLOWERING LOCUS T* (*FT*) [64, 65] and required for the maintenance of an indeterminate inflorescence meristem identity and the regulation of flowering time in *A. thaliana* and maize [63–66]. The line with the Dent candidate carried *Zmm22* (*MADS-transcription factor 69*, *Mads69*, *GRMZM2G171650*), which was recently reported to be associated with variation in flowering time in maize [67] and is considered a candidate for maize domestication and/or improvement [68, 69].

Overall, our findings in the introgression library support the relevance of the investigated genomic regions and their associated candidates for promoting flowering time and confirm the quantitative nature of flowering time in maize, determined by many genes with small effects [26] and only few genes with larger effects. Here, the effects of *Zcn1* and *Zmm22* were stronger than reported previously, which may be attributed to a stronger substitution effect when replacing a Dent haplotype with a Flint haplotype. We will target potential expression differences of flowering time candidates in the Dent–Flint introgression library in future studies to characterize possible differences in the regulation of the flowering network between germplasm pools adapted to different environments.

Differential selection on components of the flowering network within temperate maize

We investigated the 30 flowering time candidates with respect to their assignment to endogenous pathways and pathways regulated by environmental factors within the flowering network to determine if different components of the flowering network were under selective pressure in Dent and Flint, respectively. Within the flowering network, Flint candidates were involved predominantly in endogenous signaling, hormone-dependent, and developmental processes (10 of 12 candidates, 83.3 %), whereas the Dent candidates indicated a prevalence for response to environmental factors like light and photoperiod (12 of 18 candidates, 66.7 %; Fig. 3a, Additional file 4: Table S5). As described above, Flint candidates included the well-characterized *Rap2*/*Vgt1* locus and *Zcn1*. Furthermore, we found the *Squamosa promoter binding protein-transcription factor 25* (*Sbp25*, *GRMZM2G414805*) and *Gnarley1* (*Gn1*; also *Homeobox protein KNOTTED1-like 4*, *Knox4*, *GRMZM2G452178*) that are associated with aging and hormone-dependent pathways (Additional file 4: Table S5). *Gn1* is likely to act upstream of the "green revolution" gene encoding gibberellin 20-oxidase [70] and to regulate *Gibberellin 2-oxidase 1* expression in maize, thus influencing vegetative to reproductive phase transition, pollen tube growth, and stem elongation by changing the availability of gibberellin [71]. *Gibberellin 2-oxidase 1* is additionally regulated by *Knotted1* (*Kn1*, *GRMZM2G017087*) which was identified as a Dent candidate gene [71]. Another well-characterized Dent candidate is *Constans1* (*Conz1*, *GRMZM2G405368*), which is a putative ortholog of the photoperiod genes *CONSTANS* from *A. thaliana* and *Heading date1* in rice [72]. To the best of our knowledge, 20 of the 30 detected flowering time candidates have not yet been functionally characterized in the context of maize flowering time, but were associated with

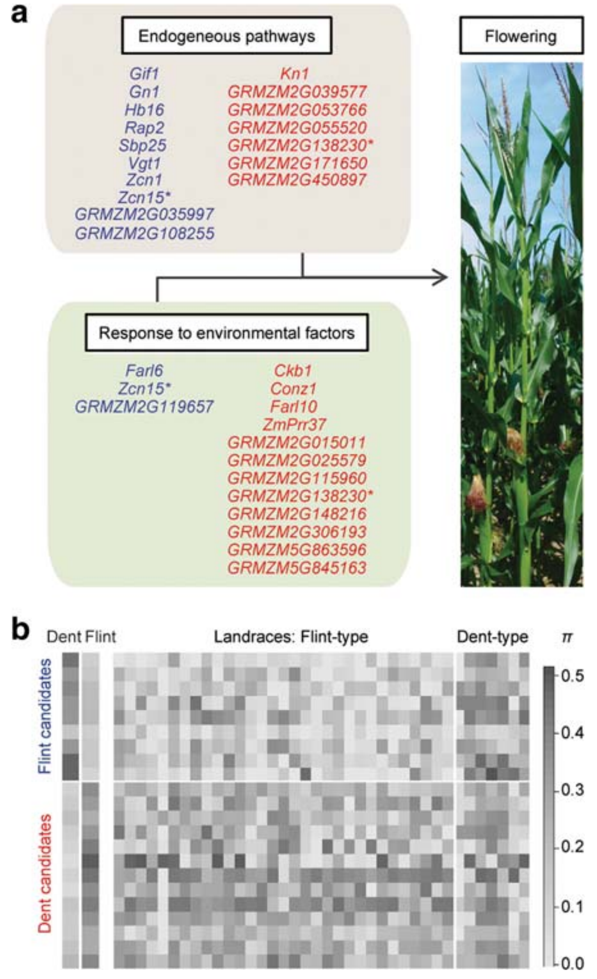

Fig. 3 Selection candidates of the maize flowering network and their nucleotide diversity in 136 elite lines and 38 European landraces. **a** Candidates associated with the maize flowering network identified under selective pressure in 70 Dent (*red*) and 66 Flint (*blue*) lines based on genotyping data. Candidates are grouped according to their putative function in endogenous pathways and pathways regulated by environmental factors. For details about the candidate genes and their classification, see Additional file 4: Table S5. Ambiguous assignments according to GO annotations and literature are indicated by *. **b** Nucleotide diversity π of nine Flint (*blue*) and 13 Dent (*red*) flowering time candidate genes for 136 temperate elite lines as well as 31 Flint-type and seven Dent-type European landraces. Mean values for each gene were calculated for the panels of Dent and Flint elite lines (*left*) and for each of the 38 landraces (*right*). For details about candidate genes, gene-wise π values, and order of landraces, see Additional file 6: Table S8

the flowering network based on GO terms or reports in other species such as *A. thaliana* and rice (Additional file 4: Table S5). Thus, our study revealed candidates that warrant further investigation of their functional relevance in maize flowering time. Based on the observed allele frequency differences of the candidate genes within the 136 elite lines and with respect to their function in maize or, for example, *A. thaliana*, we hypothesize that different components of the flowering network were under selective pressure in Dent and Flint. The Flint-specific haplotypes of these genes might constitute a promising source for the adaptation of maize germplasm pools to shorter vegetation periods.

Diversity of flowering time candidates in elite lines and European landraces

As most of the European Flint inbred lines are assumed to be derived from few landraces [44], we compared the diversity and the allelic composition of 22 flowering time candidates (13 Dent and 9 Flint candidates tagged by at least five SNPs) between the elite lines and a unique panel of 38 European landraces (Additional file 1: Table S7). For each landrace, 22 to 24 plants were genotyped at high density with the Axiom® Maize Genotyping Array [21]. The majority of the landraces (N = 31) had Flint-type kernels. These landraces exhibited lower levels of diversity in the Flint flowering time candidates (gene-wise average: $\pi = 0.130$) compared to the Dent flowering time candidates ($\pi = 0.243$), thus confirming the pattern found in the Flint elite lines (Fig. 3b, Additional file 6: Table S8). We further investigated the level of differentiation between Flint-type landraces and Flint elite lines and observed low levels of F_{ST} for the Flint flowering time candidates ($F_{ST} = 0.060$; Fig. 4a, FT) with ten landraces from France, Germany, and Spain displaying values even smaller than 0.050 (Additional file 7: Table S9). These low values of differentiation suggested a major contribution of the Flint-type landraces to the flowering time candidate gene diversity observed in the Flint elite lines. This hypothesis was corroborated by the finding that the entire set of Flint candidate genes also revealed significantly lower levels of differentiation compared to all other genes, which were not under differential selection between Dent and Flint elite lines ($F_{ST} = 0.072$ versus 0.095, p value = 6.0e-04; Fig. 4a, A versus C).

Consistent with the hypothesis that Flint elite lines and Flint-type landraces have a common history, significantly higher levels of differentiation were observed for Dent candidate genes compared to all remaining genes ($F_{ST} = 0.111$ versus 0.095, p value = 0.017; Fig. 4b, A versus C). Together, these findings indicated that the reduced diversity observed for Flint candidate genes in Flint elite lines was already present in a broad panel of European landraces and that the candidate gene diversity of the Flint elite lines originate from a limited number of Flint-type landraces used for elite line development in some historically important breeding centers [44].

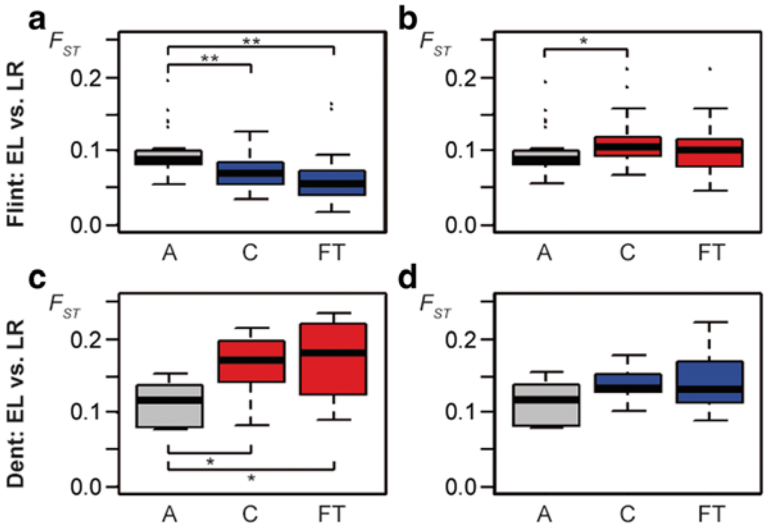

Fig. 4 Differentiation between elite lines (EL) and landraces (LR) for candidate genes. The upper panel shows the differentiation (F_{ST}) between 66 Flint elite lines and 31 Flint-type landraces for (**a**) Flint (*blue*) and (**b**) Dent (*red*) candidate gene sets. The lower panel depicts the differentiation between 70 Dent elite lines and seven Dent-type landraces for (**c**) Dent (*red*) and (**d**) Flint (*blue*) candidate gene sets. The boxplots show F_{ST} values for all (A; *gray*) genes except the candidates, the candidate (C) genes, and for the subset of candidates associated with flowering time (FT). *Boxplots* show the upper and lower quartile, median (*horizontal bar*), and whiskers (*vertical bars*) of the F_{ST} values. Points above and below the whiskers indicate values ± 1.5 times the interquartile range. Significance of two-sided Wilcoxon rank sum tests with $p < 0.05$ are indicated by * and with $p < 0.001$ by **. For details see Additional file 7: Table S9

The remaining seven landraces displayed at least partially Dent-type kernels. These landraces revealed high levels of diversity for Dent and Flint flowering time candidates ($\pi = 0.225$ and 0.260, respectively; Fig. 3b, Additional file 6: Table S8) and showed a high level of differentiation with Dent elite lines for the Dent flowering time candidates ($F_{ST} = 0.170$; Fig. 4c, FT). The same pattern was found in the analysis of the entire Dent candidate gene set, which revealed significantly higher levels of differentiation compared to all remaining genes ($F_{ST} = 0.164$ versus 0.111, p value = 0.026; Fig. 4c, A versus C), but no significant difference for Flint candidates compared to all remaining genes ($F_{ST} = 0.138$ versus 0.112, p value = 0.209; Fig. 4d, A versus C; Additional file 7: Table S9). These results indicated that the European Dent-type landraces exhibit a different allelic composition in the Dent candidates compared to the Dent elite lines and did most likely not contribute to the Dent elite material under study.

Selection on upstream and genic regions of the candidates

To examine how specific elements of the genic regions contributed to the differentiation between Dent and Flint, we compared levels of differentiation for 5 kb and 500 bp upstream regions, genic regions, and exons between the candidate gene sets and all remaining genes. To increase the resolution of our analyses, we investigated the candidate gene sets based on whole-genome sequence data of 40 temperate elite lines (21 Dent and 19 Flint) [21, 22], which were part of the panel of 136 elite lines genotyped with the 600 k array with the exception of three lines (Additional file 1: Table S1). Based on 13,246,294 bi-allelic SNPs, we observed a significant reduction of mean π and TD in 727 Dent and 403 Flint candidate genes tagged by at least five SNPs (of in total 876 and 545 candidates, respectively) compared to 31,163 remaining genes (p value < 2.2e-16; Additional file 1: Figure S5 and Additional file 1: Table S4). F_{ST} values calculated between Dent and Flint were significantly higher for candidate gene sets compared to all remaining genes for 5 kb and 500 bp upstream as well as genic and exonic regions. Together, these findings supported the results obtained from the selection screens in the panel of 136 temperate inbred lines genotyped with the 600 k array.

Previous studies in maize suggested an important role of the divergence of regulatory elements in the context of domestication [73–75]. In our study, distributions of F_{ST} values were comparable for 5 kb and 500 bp upstream as well as genic and exonic regions in each of the two candidate gene sets (Additional file 1: Figure S6 and Additional file 1: Table S4). However, the power to resolve whether selection acted differentially in upstream and genic regions was probably limited by the high level of linkage disequilibrium observed in temperate Dent and Flint lines [21]. The outcome of ongoing large-scale whole genome and transcriptome sequencing will allow the investigation of the impact of selection on the regulation of gene activity in the two pools and its consequence for the genomic differentiation between Dent and Flint.

Conclusions

In this study, we report genomic differentiation between two major temperate maize germplasm pools, Dent and Flint. By comparing a representative panel of Dent and Flint elite lines, we identified candidate genes under differential selective pressure in Dent and Flint. The significant enrichment in derived allele frequencies for these genes provided strong indication that the candidate regions represented selective sweeps. Candidate genes associated with agronomic traits known to differ between Dent and Flint could be identified. Most of the detected flowering time candidates have not yet been functionally characterized in maize. Investigating the effect of the flowering time candidates in a Dent–Flint introgression library, we found that Flint haplotypes of these candidates promoted earlier flowering. Within the flowering network of maize, a Flint-specific enrichment of genes associated with endogenous signaling, hormone-dependent pathways, and developmental processes was discovered in contrast to Dent, where selection seemed to act predominantly on genes involved in the response to environmental factors. Low levels of differentiation of Flint flowering time candidate genes between European Flint elite lines and European landraces indicated that the allelic composition of the elite lines was comparable to those of the Flint-type landraces and suggested a major contribution of landraces from France, Germany, and Spain to the candidate gene diversity in the Flint elite lines. Our findings highlight the role of genomic regions that have undergone intense selection and contributed to the differentiation of temperate Dent and Flint with likely effects on different agronomic traits. The identification of pool-specific selection signatures enabled insights into different patterns of diversity between temperate Dent and Flint and provides new targets for future functional analyses and crop improvement.

Methods

Plant material and genotyping of elite lines and landraces

The 136 elite inbred lines (Additional file 1: Table S1) were selected to represent the genetic diversity of European and American temperate maize and were genotyped with the 600 k Affymetrix® Axiom® Maize Array described in [21]. Landraces were selected to reflect the genetic and phenotypic diversity within Central and Western Europe

(Additional file 1: Table S7) and were represented by 22 to 24 plants each. All 906 landrace individuals were genotyped using the 600 k array and the genotype cluster model file (http://www.affymetrix.com/catalog/prod820010/AFFY/Axiom%26%23174%3B-Maize-Genotyping-Array).

If not denoted otherwise, analyses were performed using R version 3.0.1 [76]. SNP positions were assigned to the reference sequence B73 v2 [22] for all datasets. Analyses of the elite line panel were based on 566,961 best quality SNPs [21] with heterozygous calls masked as missing. Indels, unmapped markers, SNPs with ≥ 10 % missing values in the elite line panel [21], markers designed to specifically differentiate between two Dent lines [77], and monomorphic SNPs were excluded, resulting in 547,412 SNPs for analyses of temperate elite lines. Analyses including the landraces were based on a subset of 486,208 SNPs.

Population structure analysis and estimation of historical recombination rates

For population structure analyses of the genotyped panel of 136 elite lines, missing data were imputed using Beagle [78] version 3.3.1 via the R package "synbreed" [79] version 0.10-3. A linkage disequilibrium pruning step was performed applying an r^2 threshold of 0.8 followed by the estimation of ancestry using ADMIXTURE [80] version 1.23. To obtain an estimate of the historical recombination rates in Dent and Flint based on the genotype data of the elite line panel, the four-gamete test [51] was calculated. This test gives a conservative estimate (i.e. the minimum number) of recombination events in the history of a sample. Values for the pairwise tests of neighboring SNPs were averaged over 1000 sites and the mean of each 1000 site bin was plotted. Reported were recombination events per Mb and regions of low recombination rates were defined as regions exhibiting rates within the 10 % quantile per chromosome.

Screens for selection signatures

Nucleotide diversity π [9] and Tajima's D (TD) [10] were calculated for each panel of inbred lines using a customized script. The fixation index F_{ST} [45] was calculated across the two panels using PLINK [81] version v1.90b2m. Metrics were calculated per SNP and averaged over windows of 40 SNPs (sliding by 10 % and corresponding to an average physical distance of 10 kb), using the R package "zoo" [82]. The grid-based composite likelihood ratio (CLR) test was calculated for each panel as implemented in SweepFinder [11]. For CLR, the size of the non-overlapping grids was 150 kb, which is the same magnitude as the maximal distance between two SNPs with $r^2 > 0.2$ in the elite line panel [21]. Windows exhibiting values within the 10 % quantile (π, TD) and the 90 % quantile (F_{ST}, CLR) were submitted for candidate gene analysis based on the B73 v2 [22] annotation, version 5b60 (ftp://ftp.gramene.org/pub/gramene/maizesequence.org/release-5b/filtered-set/), containing 39,656 gene models. GO terms were tested for enrichment using the resources developed by [83] based on maize gene IDs by applying a hypergeometric test with a Benjamini–Yekutieli correction ([84]; FDR = 0.05) to account for multiple tests. Pathway analysis was performed using MapMan version 3.5.1 [53] based on the mapping of the first transcript of each gene to the file Zm_B73_5b_FGS_cds_2012.m02 downloaded from the MapMan webpage (http://mapman.gabipd.org/web/guest/mapmanstore). For visualization, arbitrary values of −4.5 and 4.5 were assigned to Dent and Flint candidate gene transcripts, respectively.

Whole-genome sequence datasets

For Fay and Wu's normalized H [52], the maize-sorghum genome alignment (http://pipeline.lbl.gov/downloads.shtml) was parsed with a custom Perl script to obtain the nucleotide in sorghum representing the ancestral maize allele for 298,388 SNPs of the genotyped panel of elite lines. Whole-genome sequence data for 30 elite lines were used to call SNPs and small indels by employing an integrative analysis pipeline [21] and filtering for high mapping (MQ ≥ 30) and genotyping quality (GQ > 5), major allele frequency ≥ 90 %, a minimal distance of 3 bp between adjacent SNPs, and a minimal and maximal coverage by three and 60 MQ30 reads for lines sequenced to medium coverage, respectively. The latter criterion was adjusted for four deep sequenced lines requiring ten and 300 MQ30 reads, respectively. We combined the obtained marker set with whole-genome sequence data from ten temperate inbred Dent and Flint lines from the maize HapMap2 project [17, 22] resulting in a VCF file including 13,246,294 bi-allelic SNPs. We filtered for SNPs with ≤ 50 % missing values across the 40 lines (F_{ST}) and within germplasm pools (π and TD). The combined VCF file was converted to hapmap format with a customized Perl script, and π and TD per gene were obtained with Variscan version 2 [85] with runmode "12".

For gene-wise calculations based on the genotyped panels of 136 elite lines and 38 European landraces (normalized H, F_{ST}) or on whole-genome sequence data of 40 elite lines (π and TD), the genic region of the longest protein-coding transcript, including 5 kb upstream, was used and metrics were calculated if at least 5 SNPs were available for analysis. For a separate analysis of exonic, genic, 500 bp, and 5 kb upstream regions, F_{ST} was determined in case of at least 5 SNPs per region based on whole-genome sequence data using vcftools v0.1.11 [86]. Two-sided Wilcoxon rank sum tests [87] were performed on gene-wise metrics to test for differences between candidate genes and remaining genes within pools.

Introgression library

To represent the Flint genome as introgression segments in a Dent genetic background, a European Dent inbred line was crossed with a European Flint inbred line, followed by backcrossing, marker-assisted selection, and several rounds of selfing [88]. The Dent parental line originated from south-eastern Europe and exhibits a high general combining ability for kernel yield. The Flint parental line was selected for high general combining ability for biomass yield and good performance in Central European climates. The introgression library used in the present study comprised 535 lines, with 97 lines carrying a single segment of the Flint parent. To estimate the length of individual donor genome fragments, the distance between markers on the respective donor genome fragment plus half the distance to the adjacent marker flanking the donor genome fragment on either side of the fragment was calculated. For the introgression lines, the two parental lines, and a check, male and female flowering times were obtained from two field experiments conducted in 2014 at the German trial locations Roggenstein (N 48°11′13.24″, E 11°19′50.86″, 517 m AMSL, average temperature in 2014: 9.8 °C) and Freising (N 48°24′11.62″, E 11°43′21.99″, 480 m AMSL, average temperature in 2014: 9.7 °C). Each experiment was laid out as an α–lattice design with two replications, except for parental lines and the check that were repeated three and five times, respectively. Male and female flowering time was recorded as days after sowing until 50 % of plants per plot exhibited emerged anthers and silks, respectively. Adjusted means for flowering time were calculated using Plabstat [89]. The difference in adjusted means between the 97 single-segment introgression lines and the Dent parental line was tested at a significance level of $\alpha = 0.05$ and $\alpha = 0.05/97$ to correct for multiple testing. Adjusted means of flowering times for seven lines carrying Flint haplotypes of Flint flowering time candidates, nine lines carrying Flint haplotypes of Dent flowering time candidates, and the 75 lines not carrying a flowering time candidate gene were tested for significant differences by calculating Student's *t*-test.

Additional files

Additional file 1: Figure S2. Variation of historical recombination rates along the ten maize chromosomes. **Figure S3.** Gene ontology (GO) terms assigned to categories of biological processes for the candidate gene sets. **Figure S4.** Pathway analysis for the Dent and Flint candidate gene sets using MapMan [53]. **Figure S5.** Gene-wise π (*left*) and TD (*right*) values based on whole-genome sequence data of 21 temperate Dent and 19 temperate Flint lines. **Figure S6.** Distributions of F_{ST} values for different genomic regions based on whole-genome sequence data of 40 elite lines. **Table S1.** Elite lines under study with germplasm pool assignment and data source. **Table S2.** Window-based statistics and thresholds according to the selected quantile based on the panel of 136 temperate inbred lines genotyped with the 600 k array. **Table S4.** Gene-wise statistics for all genes and for Dent and Flint candidate genes. **Table S7.** Landraces under study with their geographic origins. (PDF 1175 kb)

Additional file 2: Figure S1. Metrics of the selection screens for 136 temperate inbred lines along the ten maize chromosomes based on genotyping data. (PDF 4242 kb)

Additional file 3: Table S3. Gene-wise statistics for 1407 candidate genes identified in the 70 Dent (N = 876) and 66 Flint (N = 545) lines based on genotyping data. (XLSX 1238 kb)

Additional file 4: Table S5. Pathway assignment and information about candidates associated with the flowering network in maize. (XLSX 50 kb)

Additional file 5: Table S6. Genomic composition and flowering dates of 97 single-segment introgression library lines. (XLSX 114 kb)

Additional file 6: Table S8. Gene-wise nucleotide diversity π and F_{ST} for flowering time candidate genes with at least 5 SNPs in the genic and 5 kb upstream region. (XLSX 24 kb)

Additional file 7: Table S9. Gene-wise level of differentiation of allele frequencies between landraces and Dent (left) and Flint (right) elite lines based on genotyping data. (XLSX 27 kb)

Acknowledgements
We are grateful to Angel Alvarez (CSIC, Estación Experimental de Aula Dei, Zaragoza, Spain), Alain Charcosset (INRA, UMR de Génétique Quantitative et Evolution, Gif-Sur-Yvette, France), Barbara Eder and Joachim Eder (Bavarian State Research Center, Institute for Crop Science and Plant Breeding, Freising, Germany), Albrecht E. Melchinger (University of Hohenheim, Institute of Plant Breeding, Seed Science and Population Genetics, Stuttgart, Germany), Mark J. Millard (USDA-ARS, Ames, USA), Jesús Moreno González (CIAM, A Coruña, Spain), Rita Redaelli (CRA-MAC Maize Research Unit, Bergamo, Italy), Pedro Revilla (CSIC, Pontevedra, Spain), and Anne Zanetto (INRA, UE de Melgueil DiaScope, Mauguio, France) for providing elite line and/or landrace seeds. The original introgression library was kindly provided by Milena Ouzunova and Thomas Presterl (KWS SAAT SE, Einbeck, Germany). We thank Manfred Schönleben for his help with field trials and their analysis and Stefan Schwertfirm for his technical assistance, as well as Ruedi Fries (Technical University of Munich, Animal Breeding, Freising, Germany) for providing genotyping facilities. Furthermore, we thank our colleagues from the University of Hohenheim for kindly providing pictures of maize cobs and Joachim Hermisson (University of Vienna, Mathematics and BioSciences Group, Vienna, Austria) for fruitful discussions and helpful remarks during a research stay of SU in Vienna. We thank three anonymous reviewers for their valuable comments and suggestions.

Funding
This study was funded by the Federal Ministry of Education and Research (BMBF, Germany) within the AgroClustEr Synbreed - Synergistic plant and animal breeding (grant 0315528) and by the Spanish Ministry of Economy and Competitiveness (project RF2011-00022-C02-01 "Regeneration and rationalization of the maize landraces from the Iberian Peninsula"). Bernardo Ordas acknowledges a grant from the program "Ramón y Cajal" of the Spanish Ministry of Economy and Competitiveness.

Availability of data and materials
Genotype calls for elite lines and landraces are available at NCBI dbSNP upon Build B149 under accession number ss1998449529 to ss1999065412 and ss1999065413 to ss1999681296, respectively. Whole-genome sequence data for 30 elite lines are available under accession number PRJNA260788 at the NCBI BioProject webpage and the combined and filtered VCF file of whole-genome sequence data for 30 elite lines and the ten lines from the maize HapMap2 project [17, 22] are available at NCBI dbSNP under accession number ss1999681297 to ss2012927636 (upon Build B149). All three datasets are available at https://dx.doi.org/10.6084/m9.figshare.3427040.v1.

Authors' contributions

CCS, EB, AT, and SU conceived the study and discussed the results; SU investigated genotypic and outgroup data; SU and MM analyzed genotypic data of landraces; MAS, GH, and KFXM carried out variant calling; SP extracted ancestral allele information; RP, SP, and SU performed sequence analyses; PW conducted the introgression library experiments; SU and EB investigated the introgression library data; BO contributed part of the Spanish landrace data; HP performed genotyping of elite lines and landraces; SU drafted the manuscript; EB, CCS, and AT edited the manuscript; all authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Plant Breeding, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany. [2]Section of Population Genetics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany. [3]Present Address: Institute for Crop Science and Plant Breeding, Bavarian State Research Center, 85354 Freising, Germany. [4]Plant Genome and System Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [5]Misión Biológica de Galicia, Spanish National Research Council (CSIC), 36080 Pontevedra, Spain. [6]Animal Breeding, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany.

Received: 16 December 2015 Accepted: 15 June 2016
Published online: 08 July 2016

References

1. Lobell DB, Tebaldi C. Getting caught with our plants down: the risks of a global crop yield slowdown from climate trends in the next two decades. Environ Res Lett. 2014;9:074003.
2. Smith CW, Betrán J, Runge ECA. Corn: origin, history, technology, and production. 1st ed. Hoboken, NJ: John Wiley & Sons Inc.; 2004.
3. Brown WL, Anderson E. The northern flint corn. Ann Mo Bot Gard. 1947;34:1–28.
4. Rebourg C, Chastanet M, Gouesnard B, Welcker C, Dubreuil P, Charcosset A. Maize introduction into Europe: the history reviewed in the light of molecular data. Theor Appl Genet. 2003;106:895–903.
5. Dubreuil P, Dufour P, Krejci E, Causse M, deVienne D, Gallais A, et al. Organization of RFLP diversity among inbred lines of maize representing the most significant heterotic groups. Crop Sci. 1996;36:790–9.
6. Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, et al. Intraspecific variation of recombination rate in maize. Genome Biol. 2013;14:R103.
7. Giraud H, Lehermeier C, Bauer E, Falque M, Segura V, Bauland C, et al. Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. Genetics. 2014;198:1717–34.
8. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39:197–218.
9. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105:437–60.
10. Tajima F. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.
11. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15:1566–75.
12. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832–7.
13. Voight BF, Kudaravalli S, Wen XQ, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:446–58.
14. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat Genet. 2013;45:884–90.
15. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 2009;19:826–37.
16. Qanbari S, Gianola D, Hayes B, Schenkel F, Miller S, Moore S, et al. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics. 2011;12:318.
17. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia JM, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44:808–U118.
18. Jiao YP, Zhao HN, Ren LH, Song WB, Zeng B, Guo JJ, et al. Genome-wide genetic changes during modern breeding of maize. Nat Genet. 2012;44:812–U124.
19. Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, et al. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Natl Acad Sci U S A. 2002;99:9650–5.
20. Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF, et al. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell. 2005;17:2859–72.
21. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics. 2014;15:823.
22. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012;44:803–7.
23. Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Barraud S, et al. Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. Genetics. 2006;172:2449–63.
24. Revilla P, Rodríguez VM, Ordás A, Rincent R, Charcosset A, Giauffret C, et al. Cold tolerance in two large maize inbred panels adapted to European climates. Crop Sci. 2014;54:1981–91.
25. Chardon F, Virlon B, Moreau L, Falque M, Joets J, Decousset L, et al. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. Genetics. 2004;168:2169–85.
26. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al. The genetic architecture of maize flowering time. Science. 2009;325:714–8.
27. Colasanti J, Yuan Z, Sundaresan V. The *indeterminate* gene encodes a zinc finger protein and regulates a leaf-generated signal required for the transition to flowering in maize. Cell. 1998;93:593–603.
28. Muszynski MG, Dam T, Li B, Shirbroun DM, Hou Z, Bruggemann E, et al. *Delayed flowering1* encodes a basic leucine zipper protein that mediates floral inductive signals at the shoot apex in maize. Plant Physiol. 2006;142:1523–36.
29. Salvi S, Tuberosa R, Chiapparino E, Maccaferri M, Veillet S, van Beuningen L, et al. Toward positional cloning of *Vgt1*, a QTL controlling the transition from the vegetative to the reproductive phase in maize. Plant Mol Biol. 2002;48:601–13.
30. Vladutu C, McLaughlin J, Phillips RL. Fine mapping and characterization of linked quantitative trait loci involved in the transition of the maize apical meristem from vegetative to generative structures. Genetics. 1999;153:993–1007.
31. Chen C, DeClerck G, Tian F, Spooner W, McCouch S, Buckler E. PICARA, an analytical pipeline providing probabilistic inference about a priori candidates genes underlying genome-wide association QTL in plants. PLoS One. 2012;7:e46596.
32. Danilevskaya ON, Meng X, Hou Z, Ananiev EV, Simmons CR. A genomic and expression compendium of the expanded *PEBP* gene family from maize. Plant Physiol. 2008;146:250–64.
33. Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS One. 2012;7:e43450.
34. Ducrocq S, Madur D, Veyrieras JB, Camus-Kulandaivelu L, Kloiber-Maitz M, Presterl T, et al. Key impact of *Vgt1* on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics. 2008;178:2433–7.
35. Xu J, Liu Y, Liu J, Cao M, Wang J, Lan H, et al. The genetic architecture of flowering time and photoperiod sensitivity in maize as revealed by QTL review and meta analysis. J Integr Plant Biol. 2012;54:358–373.

36. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 2013;14:R55.
37. Castelletti S, Tuberosa R, Pindo M, Salvi S. A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL Vgt1. G3 (Bethesda). 2014;4:805–12.
38. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. Genetic properties of the maize nested association mapping population. Science. 2009;325:737–40.
39. Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, et al. Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. Genetics. 2014;198:3–16.
40. Mikel MA. Availability and analysis of proprietary dent corn inbred lines with expired US plant variety protection. Crop Sci. 2006;46:2555–60.
41. Mikel MA, Dudley JW. Evolution of North American dent corn from public to proprietary germplasm. Crop Sci. 2006;46:1193–205.
42. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 2005;44:1054–64.
43. Nelson PT, Coles ND, Holland JB, Bubeck DM, Smith S, Goodman MM. Molecular characterization of maize inbreds with expired US plant variety protection. Crop Sci. 2008;48:1673–85.
44. Barrière Y, Alber D, Dolstra O, Lapierre C, Motto M, Ordas A, et al. Past and prospects of forage maize breeding in Europe: II. History, germplasm evolution and correlative agronomic changes. Maydica. 2006;51:435–49.
45. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38:1358–70.
46. Parat F, Schwertfirm G, Rudolph U, Miedaner T, Korzun V, Bauer E, et al. Geography and end use drive the diversification of worldwide winter rye populations. Mol Ecol. 2016;25:500–14.
47. Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, et al. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genet. 2014;10:e1004148.
48. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet. 2012;44:212–6.
49. Stephan W. Genetic hitchhiking versus background selection: the controversy and its implications. Philos Trans R Soc Lond B Biol Sci. 2010;365:1245–53.
50. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134:1289–303.
51. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985;111:147–64.
52. Zeng K, Fu YX, Shi SH, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006;174:1431–9.
53. Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, et al. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 2004;37:914–39.
54. Zhou M, Zhang Q, Wang C, Chen L, Sun Z, Zhu X, et al. Characterization of genes involved in isoprenoid diphosphate biosynthesis in maize. J Plant Growth Regul. 2015;34:294–308.
55. van Schie CC, Ament K, Schmidt A, Lange T, Haring MA, Schuurink RC. Geranyl diphosphate synthase is required for biosynthesis of gibberellins. Plant J. 2007;52:752–62.
56. Degen T, Dillmann C, Marion-Poll F, Turlings TC. High genetic variability of herbivore-induced volatile emission within a broad range of maize inbred lines. Plant Physiol. 2004;135:1928–38.
57. Kollner TG, Held M, Lenk C, Hiltpold I, Turlings TC, Gershenzon J, et al. A maize (E)-beta-caryophyllene synthase implicated in indirect defense responses against herbivores is not expressed in most American maize varieties. Plant Cell. 2008;20:482–94.
58. Trzcinska-Danielewicz J, Bilska A, Fronk J, Zielenkiewicz P, Jarochowska E, Roszczyk M, et al. Global analysis of gene expression in maize leaves treated with low temperature I. Moderate chilling (14 °C). Plant Sci. 2009;177:648–58.
59. Sobkowiak A, Jonczyk M, Jarochowska E, Biecek P, Trzcinska-Danielewicz J, Leipner J, et al. Genome-wide transcriptomic analysis of response to low temperature reveals candidate genes determining divergent cold-sensitivity of maize inbred lines. Plant Mol Biol. 2014;85:317–31.
60. Yan SP, Zhang QY, Tang ZC, Su WA, Sun WN. Comparative proteomic analysis provides new insights into chilling stress responses in rice. Mol Cell Proteomics. 2006;5:484–96.
61. Hung HY, Shannon LM, Tian F, Bradbury PJ, Chen C, Flint-Garcia SA, et al. *ZmCCT* and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proc Natl Acad Sci U S A. 2012;109: E1913–21.
62. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci U S A. 2007;104:11376–81.
63. Danilevskaya ON, Meng X, Ananiev EV. Concerted modification of flowering time and inflorescence architecture by ectopic expression of *TFL1*-like genes in maize. Plant Physiol. 2010;153:238–51.
64. Kobayashi Y, Kaya H, Goto K, Iwabuchi M, Araki T. A pair of related genes with antagonistic roles in mediating flowering signals. Science. 1999;286:1960–2.
65. Koornneef M, Hanhart CJ, van der Veen JH. A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. Mol Gen Genet. 1991;229:57–66.
66. Hanano S, Goto K. *Arabidopsis TERMINAL FLOWER1* is involved in the regulation of flowering time and inflorescence development through transcriptional repression. Plant Cell. 2011;23:3172–84.
67. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 2014;26:121–35.
68. Briggs WH, McMullen MD, Gaut BS, Doebley J. Linkage mapping of domestication loci in a large maize teosinte backcross resource. Genetics. 2007;177:1915–28.
69. Zhao Q, Weber AL, McMullen MD, Guill K, Doebley J. MADS-box genes of maize: frequent targets of selection during domestication. Genet Res (Camb). 2011;93:65–75.
70. Sasaki A, Ashikari M, Ueguchi-Tanaka M, Itoh H, Nishimura A, Swapan D, et al. Green revolution: A mutant gibberellin-synthesis gene in rice - New insight into the rice variant that helped to avert famine over thirty years ago. Nature. 2002;416:701–2.
71. Bolduc N, Hake S. The maize transcription factor KNOTTED1 directly regulates the gibberellin catabolism gene *ga2ox1*. Plant Cell. 2009;21:1647–58.
72. Miller TA, Muslin EH, Dorweiler JE. A maize *CONSTANS*-like gene, *conz1*, exhibits distinct diurnal expression patterns in varied photoperiods. Planta. 2008;227:1377–88.
73. Wang RL, Stec A, Hey J, Lukens L, Doebley J. The limits of selection during maize domestication. Nature. 1999;398:236–9.
74. Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, et al. The origin of the naked grains of maize. Nature. 2005;436:714–9.
75. Lemmon ZH, Bukowski R, Sun Q, Doebley JF. The role of *cis* regulatory evolution in maize domestication. PLoS Genet. 2014;10:e1004745.
76. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation; 2013.
77. Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, et al. A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS One. 2011;6:e28334.
78. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210–23.
79. Wimmer V, Albrecht T, Auinger HJ, Schön C-C. Synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics. 2012;28:2086–7.
80. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
81. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
82. Zeileis A, Grothendieck G. zoo: S3 infrastructure for regular and irregular time series. J Stat Softw. 2005;14:1–27.
83. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38:W64–70.
84. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 2001;29:1165–88.
85. Hutter S, Vilella AJ, Rozas J. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics. 2006;7:409.

86. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

87. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1:80–3.

88. Gresset S, Westermeier P, Rademacher S, Ouzunova M, Presterl T, Westhoff P, et al. Stable carbon isotope discrimination is under genetic control in the C_4 species maize with several genomic regions influencing trait expression. Plant Physiol. 2014;164:131–43.

89. Utz HF. PLABSTAT - A computer program for statistical analysis of Plant Bred experiments. Version 3A. Stuttgart: Universität Hohenheim; 2011.

8 Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Chris-Carolin Schön. She gave me the opportunity to work on this fascinating topic within the framework of the Synbreed project. During this time, I learned a lot thanks to her excellent guidance on research and her encouragement to continuous improvement. Furthermore, I sincerely thank Prof. Dr. Aurélien Tellier for taking his time discussing questions and sharing his thoughts with me. I am grateful for his support and his invaluable advice during the last years. To my delight Prof. Dr. Arthur Korte accepted to serve on my graduate committee and I would also like to express my thanks to Prof. Dr. Brigitte Poppenberger for chairing my graduate committee.

I would like to gratefully acknowledge and sincerely thank my supervisor Dr. Eva Bauer for her constant assistance and kindly sharing her expertise with me. Additionally, I would like to express my special appreciation and gratitude to my mentor Prof. Dr. Joachim Hermisson, Dr. Andrea Betancourt, Christian Huber as well as Prof. Dr. Magnus Nordborg and his research group for all their helpful remarks, interesting discussions and their great support on population genetic questions.

Special thanks go to all my former and present colleagues at the Chair of Plant Breeding as well as Saurabh Pophaly and Michael Seidel for their continuous assistance, motivation and lots of joy throughout my PhD.

This research was funded by the Federal Ministry of Education and Research (BMBF, Germany) within the AgroClustEr Synbreed - Synergistic plant and animal breeding (grant 0315528).

Finally, I would like to thank my beloved parents. Words cannot express my feelings and my gratitude - thank you for all your unconditional support and encouragement throughout the years!