Theor Appl Genet (2014) 127:2545–2553
DOI 10.1007/s00122-014-2396-6

ORIGINAL PAPER

Identification of key ancestors of modern germplasm in a breeding program of maize

F. Technow · T. A. Schrag · W. Schipprack · A. E. Melchinger

Received: 7 June 2014 / Accepted: 28 August 2014 / Published online: 11 September 2014

Abstract
Key message **Probabilities of gene origin computed from the genomic kinships matrix can accurately identify key ancestors of modern germplasms**
Abstract Identifying the key ancestors of modern plant breeding populations can provide valuable insights into the history of a breeding program and provide reference genomes for next generation whole genome sequencing. In an animal breeding context, a method was developed that employs probabilities of gene origin, computed from the pedigree-based additive kinship matrix, for identifying key ancestors. Because reliable and complete pedigree information is often not available in plant breeding, we replaced the additive kinship matrix with the genomic kinship matrix. As a proof-of-concept, we applied this approach to simulated data sets with known ancestries. The relative contribution of the ancestral lines to later generations could be determined with high accuracy, with and without selection. Our method was subsequently used for identifying the key ancestors of the modern Dent germplasm of the public maize breeding program of the University of Hohenheim. We found that the modern germplasm can be traced back to six or seven key ancestors, with one or two of them having a disproportionately large contribution. These results largely corroborated conjectures based on early records of the breeding program. We conclude that probabilities of gene origin computed from the genomic kinships matrix can be used for identifying key ancestors in breeding programs and estimating the proportion of genes contributed by them.

Communicated by Natalia de Leon.

F. Technow (✉) · T. A. Schrag · W. Schipprack · A. E. Melchinger
Department of Applied Genetics, Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart 70599, Germany
e-mail: Frank.Technow@uni-hohenheim.de

Introduction

Key ancestors are individuals of an ancestral population with a major contribution of genes to the germplasm pool of the modern population. Their identification can provide insights into the historical development of breeding programs and can help to quantify the impact of past breeding decisions. Knowing the major contributing sources may also help in identifying underutilized germplasm sources for broadening the genetic diversity of a breeding program.

Next generation whole genome sequencing (NGS) offers a vast array of new opportunities for genomic breeding of animals and plants. While whole genome sequences become cheaper year after year, sequencing all inbred lines in a population is still far out of reach. However, using novel imputation techniques, it may suffice to sequence only a core subset of individuals and impute the genotypes of the remainder (Kong et al. 2008; Daetwyler et al. 2011). The concept of key ancestors can serve as a rationale for identification of such a core set. Provided the ancestral and modern population are not separated by too many generations, ancestral chromosome segments are still large enough to be identified in the modern population with low to intermediate marker density (Daetwyler et al. 2011; Goddard and Hayes 2009a; Kong et al. 2008). Thus, key ancestors could serve as reference genomes for imputing low-density marker data to the sequence level (Goddard and Hayes 2009a; Hayes and Bowman 2011).

The contribution of ancestors to the modern population can be quantified as marginal probabilities of gene origin

(PGO). Boichard et al. (1997) described several pedigree-based methods for computing PGO in the context of animal breeding. More recently, Goddard and Hayes (2009a) proposed a novel PGO method which is based on the expected additive kinship matrix. Pausch et al. (2013) used this method successfully for identifying informative subsets of 'Fleckvieh' cattle bulls for sequencing.

Different from the situation in animal breeding, pedigree information in plant breeding is often incomplete, ambiguous, and erroneous (Graner et al. 1994; Lübberstedt et al. 2000; El-Kassaby et al. 2011). The methods described by Boichard et al. (1997) are therefore not applicable here. Genetic relationships, whether computed from markers or pedigree, are defined relative to a hypothetical or specified base population of individuals that are assumed or declared to be unrelated (Falconer and Mackay 1996). However, setting the coancestry coefficients between individuals to zero simply because of incomplete or missing pedigree data, leads to an underestimation of true coancestry (Messmer et al. 1993). Because of the immense mathematical difficulties in developing a comprehensive theory for the joint effects of selection and random drift in finite populations (Kimura 1964), the computation of coancestry coefficients from pedigrees is based on the simplifying assumption of the absence of selection. This assumption is often clearly violated in a plant breeding context (Cox et al. 1985; Messmer et al. 1991). Owing to these shortcomings, various authors (e.g., Messmer et al. 1991, 1993; Smith et al. 1997; Van Inghelandt et al. 2010) concluded that marker-based estimates reflect the actual genetic relationships better than pedigree-based estimates. Novel genotyping platforms facilitate genotyping on a large scale at reasonable costs (Elshire et al. 2011; Ganal et al. 2011). Thus, by replacing the expected kinship matrix with a marker-based estimate (Eding and Meuwissen 2001), the method proposed by Goddard and Hayes (2009a) might be suitable for identifying key ancestors of plant breeding populations, too.

Our objectives were to (1) provide a proof-of-concept for using the genomic kinship matrix to identify key ancestors in a simulated data set with known ancestries in the presence and absence of selection and (2) demonstrate the use of this method for identifying key ancestors in populations of modern Dent germplasm from a public maize breeding program.

Materials and methods

Identification of key ancestors

Let $\Pi = \{1, 2, \ldots, N\}$ denote the subset of individuals analyzed from the modern population and $\Gamma = \{1, 2, \ldots, C\}$ a set of potential ancestors of Π from which we want to identify a set of key ancestors on the basis of the criteria defined below. Let $\boldsymbol{K}_\Gamma$ denote the $C \times C$ genomic kinship matrix of the genotypes in set Γ and $\boldsymbol{K}_{\Gamma,\Pi}$ denote the $C \times N$ genomic kinship matrix between the genotypes in Γ and Π. Then, $\boldsymbol{c} = \boldsymbol{K}_{\Gamma,\Pi}\mathbf{1}N^{-1}$ is the vector of average genomic kinships between the potential ancestors in Γ and the modern population Π. Following Goddard and Hayes (2009a), the vector of PGO of the potential ancestors can be computed as $\boldsymbol{b} = \boldsymbol{K}_\Gamma^{-1}\boldsymbol{c}$. An element b_i of vector $\boldsymbol{b}$ can be interpreted as the probability that a randomly drawn gene copy from the modern population Π originates from ancestor i, i.e., the PGO. Thus, $\mathbf{1}'\boldsymbol{b}$ gives the total probability that a gene of the modern population descends from a given set of ancestors. We used a forward selection approach to identify a subset of key ancestors, as proposed by Goddard and Hayes (2009a). Here, the individual with the largest element in $\boldsymbol{b}$ is considered as the first key ancestor. The second key ancestor is the individual which leads to the highest increase in $\mathbf{1}'\boldsymbol{b}_{n=2}$, where $\boldsymbol{b}_{n=2}$ is the vector of marginal probabilities of the first and second key ancestor. This vector was recomputed for every combination and number of ancestors considered. The selection process was repeated until $\mathbf{1}'\boldsymbol{b}_{n+1} - \mathbf{1}'\boldsymbol{b}_n < 0.015$. The threshold value of 0.015 was found to be suitable in a preliminary simulation study. However, other similarly low values could have been used as well.

Boichard et al. (1997) calculate the effective number of ancestors as $f_a = 1/\boldsymbol{b}_n'\boldsymbol{b}_n$, where n corresponds to the number of key ancestors identified. For taking into account that $\mathbf{1}'\boldsymbol{b}_n$ can be considerably smaller than 1.0, we propose to modify this formula and computed the normalized effective number of ancestors $F_a = (\mathbf{1}'\boldsymbol{b}_n)^2/\boldsymbol{b}_n'\boldsymbol{b}_n$. Computing F_a in this way ensured that the maximum of F_a is equal to n even if $\mathbf{1}'\boldsymbol{b}_n < 1$. The more unbalanced the contribution of the key ancestors to the modern population, the lower F_a.

We used the method of Eding and Meuwissen (2001) for computing the genomic kinship matrices $\boldsymbol{K}_\Gamma$ and $\boldsymbol{K}_{\Gamma,\Pi}$. This approach results in estimates that are directly interpretable as probabilities of identity by descent (IBD). First, pairwise similarity scores S_{ij}, averaged over all markers, were computed. These scores were then converted to estimates of IBD kinship (f_{ij}) by adjusting with the probability of identity in state (s)

$$f_{ij} = \frac{S_{ij} - s}{1 - s}, \tag{1}$$

where the minimum value of S_{ij} in the ancestral population Γ was used as an estimate of s.

Data simulation

We stochastically simulated an ancestral population Γ of 50 inbred lines. The genome consisted of ten chromosomes of

length 1 Morgan (M) each. We placed 500 equally spaced biallelic marker loci on each chromosome and simulated low historical linkage disequilibrium (LD) between the loci that decayed exponentially such that the observed LD was halved for every 0.01 M distance. Specifically, the expected LD, measured as r^2 between two loci with a genetic distance of t M apart, was equal to $0.10 \times 2^{-t/0.01}$. An exponential decay curve closely mirrors the decay curves observed in maize (Technow et al. 2013) and other species (Goddard and Hayes 2009b). Allele frequencies were drawn from a uniform distribution in the interval [0.35, 0.65]. The stochastic simulation of LD and allele frequencies in the ancestral population was performed with the algorithm described in Montana (2005), using a customized version of the program code of the R (R Core Team 2012) package accompanying their publication. Meiosis was simulated according to the assumptions underlying the Haldane mapping function, using the R package 'hypred' (Technow 2012).

In addition to the 500 observed biallelic markers, we placed 500 equally spaced 'tag' markers on each ancestral chromosome, with alleles unique to each ancestor. This allowed us to track ancestral chromosome segments throughout the subsequent recombination cycles and compute the true values of vector $\boldsymbol{b}$ reflecting the marginal contributions of an ancestor to the modern population Π.

Our simulation of selection in each generation followed the one conducted by Stich et al. (2007) to investigate the causes of LD in a typical Central European maize breeding program. We performed random crosses between the ancestors in Γ and generated $N = 200$ recombinant doubled haploid lines (DH) through a chromosome duplication step. Thereby, we allowed crosses to appear multiple times. The 200 DH lines formed the base generation of cycle C_1. From this generation, 25 DH lines were selected according to two scenarios, which are described below, and again 200 crosses were produced among them to generate cycle C_2. This scheme was repeated until cycle C_9, from which 200 individuals were obtained and considered as the modern population Π. Our goal was to identify the key ancestors of population Π, from the ancestral population Γ, using the methods described above.

In scenario 1 ('neutral scenario'), the 25 DH lines used as parents for generating the next cycle were always chosen at random. In scenario 2 ('selection scenario'), selection of the parents occurred on the basis of the phenotypic value. The latter was simulated as the sum of the genotypic value and a normally distributed noise variable, as described below. The variance of this noise variable was held constant over all cycles and was chosen such that the heritability $h^2 = 0.5$ in cycle C_1. The genotypic values were simulated by assigning additive effects, drawn from a standard normal distribution, to a random subset of 1,000 of the markers. The true marginal contribution of an ancestor in Γ to the modern population Π was computed as the proportion of 'tag' marker genotypes that characterized this ancestor. In total, 100 data sets were generated by repeating the full simulation process.

Dent populations

Modern populations and potential ancestors

The modern elite germplasm was represented by (1) 36 Dent lines with pure Stiff-Stalk-Synthetic background, subsequently referred to as SSS-Dent lines and (2) 136 further Dent lines comprising various sources of Dent but mostly Iodent, subsequently denoted as IOD-Dent lines. All lines have been developed from biparental crosses or synthetic multi-parent populations by recurrent selfing, accompanied by selection, for at least six selfing generations. These inbred lines represent the current elite material of the public maize breeding program of the University of Hohenheim. They have been selected in multiple steps for (a) line per se performance in all selfing generations and in parallel (b) testcross performance of S_2 to S_5 lines with one or two testers in one or two years at three to four locations for traits relevant for grain maize production in Central Europe. Finally, the best 10 to 15 lines in each year were used as parents of factorial crosses evaluated in 6 to 8 environments. A detailed overview about these lines and the factorials in which they were evaluated, was presented by Technow et al. (2014).

Based on pedigree records of the modern breeding lines, which were often incomplete or included materials that were unknown or no longer accessible, we identified 13 historical Dent lines developed mainly in the 1950s and 1960s as potential key ancestors of the modern elite germplasm developed by the University of Hohenheim (Table 1).

Genomic data

The modern and historical inbred lines were genotyped with the Illumina MaizeSNP50 Bead Chip (Ganal et al. 2011). All markers with more than 5 % missing or more than 5 % heterozygous marker genotypes were removed. Remaining missing (0.9 %) or heterozygous (0.8 %) marker genotypes were replaced with the allele, which had the highest frequency in historical and modern lines. A total of 40,982 markers were subsequently available for further analysis.

Key ancestors were identified separately for the modern SSS-Dent and IOD-Dent breeding populations using the method described above. For identifying key ancestors of SSS-Dent, only markers with minor allele frequency (MAF) above 0.025 in the combined set of historical and

Table 1 Description of historical inbred lines considered as potential ancestors of modern Dent germplasm used in the maize breeding program of the University of Hohenheim

Background	Historical line	Pedigree/source	Origin	References
SSS	A632	[(Mt42 × B14) B14 (3)]	Minnesota	Gerdes et al. (1993)
	B37	Iowa Stiff-Stalk-Synthetic	Iowa	Gerdes et al. (1993)
	B73	Iowa Stiff-Stalk-Synthetic	Iowa	Gerdes et al. (1993)
Lancaster	Mo17	C.I. 187-2 × C103	Missouri	Gerdes et al. (1993)
	Oh43	W8 × Oh40B	Ohio	Gerdes et al. (1993)
NSS	W401	[(33 × Wisconsin No. 25) × 67C]	Wisconsin	Gerdes et al. (1993)
	W59E	[(WM13 × W.Va352) × (W9 × A49)]	Wisconsin	Gerdes et al. (1993)
	W41A	WH × WJ	Wisconsin	Gerdes et al. (1993)
	W153R	[(Ia153 × W8) Ia153]	Wisconsin	Gerdes et al. (1993)
Early Butler	Co109	Early Butler	Ottawa	Gerdes et al. (1993)
Unknown	Co125	Unknown	Ottawa	Messmer et al. (1993)
	Wf9	Unknown	Indiana	Gethi et al. (2002)
Iodent	IOD-663	Unknown	Unknown	W. Schipprack (personal communication)

SSS-Dent lines were used for computing the kinship matrices $\boldsymbol{K}_{\Gamma}$ and $\boldsymbol{K}_{\Gamma,\Pi}$. In the case of IOD-Dent, only markers with MAF above 0.025 in the combined set of historical and IOD-Dent lines were used.

Modified Roger's distances (MRD) (Wright 1978) between all pairs of lines were computed from the 31,798 markers polymorphic in the combined set of historical and modern lines with MAF >0.025. These values were subsequently used for performing a principal coordinate analysis (PCoA) according to Gower (1966).

Results

Simulated data

Genetic characteristics

Averaged over replications, the average MAF was 0.26 in the neutral scenario and 0.20 in the selection scenario. In the neutral scenario, an average of 27.7 ancestors had a true marginal contribution greater than zero, in the selection scenario this value was 22.3. The three ancestors with the highest contribution contributed a total proportion of 0.30 in the neutral scenario and 0.42 in the selection scenario.

Identification of key ancestors

On average, 14.7 key ancestors with a true total marginal contribution of 0.83 and an estimated total marginal contribution of 0.94 were identified in the neutral scenario. In the selection scenario, we identified 12.5 key ancestors on average. Here, the true total and estimated total contributions were 0.88 and 0.97, respectively. The curve of true cumulative contributions for the selection scenario was very steep for the first few ancestors, but flattened considerably as more ancestors were added (Fig. 1). The same trend held true for the neutral scenario, but was less pronounced. On average, the normalized effective number of ancestors F_a amounted to 12.9 in the neutral scenario and 9.8 in the selection scenario.

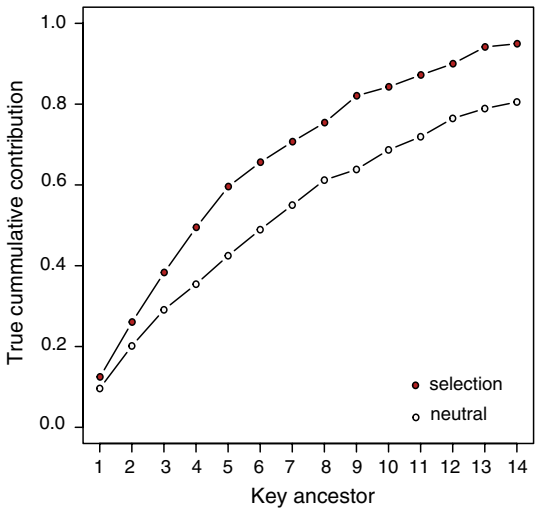

Fig. 1 True cumulative contributions of key ancestors, ranked from highest to lowest contributing, in representative examples of the neutral (*empty circles*) and selection (*full circles*) scenario

The Spearman rank correlation between true and estimated marginal contributions of identified key ancestors averaged 0.87 and 0.90 in the neutral and selection scenario, respectively (Fig. 2). We repeated our simulations

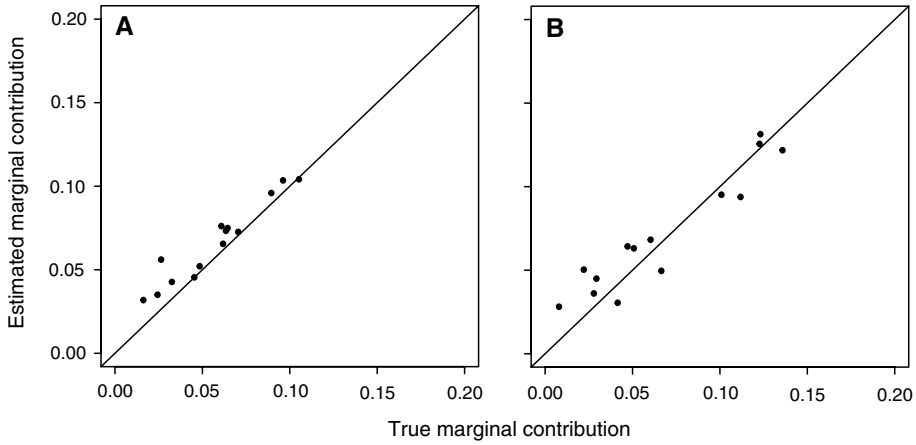

Fig. 2 Scatter plot of true and estimated marginal contributions of key ancestors in representative examples of (**a**) the neutral and (**b**) the selection scenario

also with an ancestral population Γ with a reduced size of 15, which more closely matched the number of potential key ancestors of the Dent germplasms in the breeding program analyzed. Here, the Spearman rank correlation between estimated and true marginal contributions of the identified key ancestors was also close to 0.90 (data not shown).

Experimental data

Principal coordinate analysis

The average MAF within the historical Dent lines was 0.24. For the modern IOD-Dent and SSS-Dent germplasm, the average MAF was 0.16 and 0.15, respectively. The first two principal coordinates (PC) explained 48.7 and 13.8 % of the molecular variation among all, respectively. PC1 clearly separated the two modern Dent germplasm groups (Fig. 3). It also separated the historical Dent line IOD-0663 from the other historical Dent lines, with the modern IOD-Dent lines scattered between them. The historical Dent lines A632, B73 and Mo17 were slightly distant from the other historical Dent lines and located among the modern SSS-Dent lines on the axis of PC1.

Identification of key ancestors

For the modern SSS-Dent population, the key ancestors and their marginal contribution were: A632 (0.27), Co125 (0.17), W59E (0.11), Oh43 (0.10), Wf9 (0.10), Mo17 (0.06) and B73 (0.05), with a total contribution of 0.87 (Fig. 4a). For SSS-Dent, F_a was equal to 5.33.

The identified key ancestors and their marginal contribution for the IOD-Dent population were: IOD-0663 (0.49), Co125 (0.13), A632 (0.12), W59E (0.08), Mo17 (0.05), and B73 (0.04) (Fig. 4b). The combined contribution of all six ancestors was 0.90, of which line IOD-0663 contributed 54 %. The estimate of F_a was equal to 2.93.

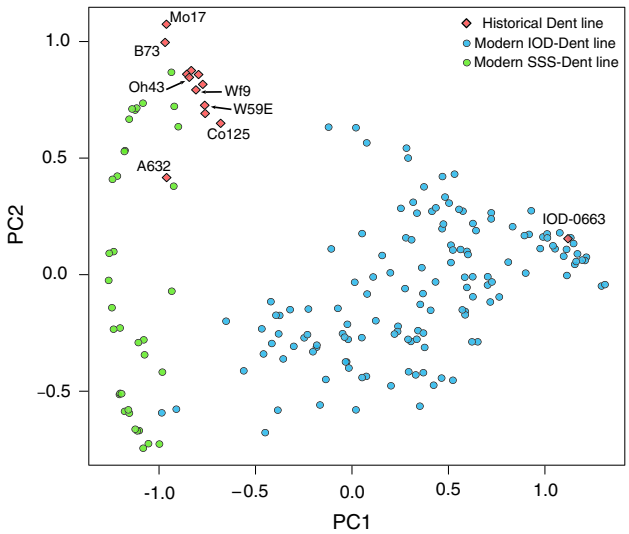

Fig. 3 *Scatter plot* of first principal coordinate (PC1) (48.7 % variance explained) versus second principal coordinate PC2 (13.8 % variance explained) for the combined set of historical and modern Dent and populations. Historical lines identified as key ancestors of either SSS-Dent or IOD-Dent are identified by name

Discussion

Simulated data

The main goal of the analyses of simulated data sets was to provide a proof-of-concept that the genomic kinship matrix can be used for computing PGO and for identifying key ancestors for populations of inbred lines using the method of Goddard and Hayes (2009a). The high rank correlation between estimated and true contribution showed that the differences in the importance of the ancestral lines were captured accurately. As revealed by the 'tag' markers and expected from theory, applying directional selection decreased the number of ancestral lines contributing to the modern population, compared to the scenario without

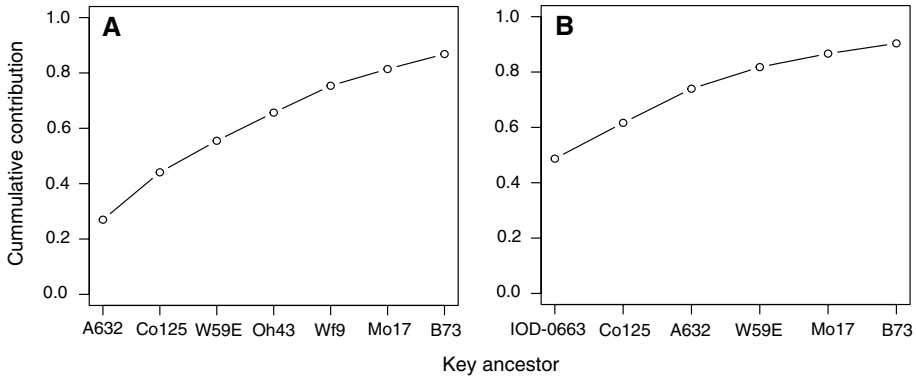

Fig. 4 Cumulative estimated contributions of key Dent ancestor lines, for (**a**) modern SSS-Dent and (**b**) IOD-Dent populations, ranked from highest to lowest contributing ancestor

selection. This explains why we identified fewer key ancestors in the selection scenario than in the neutral scenario. The normalized effective number of ancestors F_a was closer to the number of identified key ancestors n in the neutral scenario than it was in the selection scenario. This indicates that the contribution of key ancestors to the modern population was more unbalanced in the selection scenario, which reflects that the genome of superior ancestors was enriched in the modern population. Thus, the observed differences in n and F_a between the two scenarios demonstrated that the method was sensitive enough to reflect the differences between the historical development of the modern populations in the neutral and selection scenario.

The number of identified key ancestors was considerably smaller than the number of ancestral lines with a non-zero true contribution to the modern population. However, the identified key ancestors accounted for a true total contribution of more than 80 % in both scenarios. This shows that the method succeeded in separating ancestors with a minor contribution from those with a major contribution, i.e., the key ancestors. The estimated total contribution of the selected key ancestors overestimated their true total contribution. An explanation is that an ancestor for which the marginal contribution is overestimated has a higher chance to be included into the set of key ancestors than an ancestor for which the marginal contribution is not overestimated.

Key ancestors of modern Dent germplasm

In the US Cornbelt, hybrids are generally produced between lines from different heterotic groups such as Stiff-Stalk-Synthetic, Lancaster and Iodent (Mikel and Dudley 2006). In contrast, the predominant heterotic pattern in Central Europe are hybrids of type Dent × Flint (Messmer et al. 1993). Therefore, the various Dent subgroups must not necessarily be kept separate and intermating between them for the breeding of Dent lines is not unusual. In fact, the pedigree records of the Dent germplasm analyzed in this study reveal that the current Dent heterotic pool of the maize breeding program of the University of Hohenheim represents a mixture of various original Dent sources, as applies to many maize breeding programs in Central Europe.

Proprietary pedigree records from the beginning of the University of Hohenheim's maize breeding program, as far as available, suggest that its IOD-Dent population was established in the 1980s from crosses between the Iodent line IOD-0663 and several SSS-Dent lines. This was confirmed by our analysis, where the most prominent key ancestor of the modern IOD-Dent population was IOD-0663, with a marginal contribution of almost 50 %. This is the contribution expected, if the IOD-Dent population was established from biparental crosses in which IOD-0663 was consistently used as a parent. The historical lines A632, Co125, W59E, Mo17 and B73 were among the key ancestors of both IOD-Dent and SSS-Dent populations. This overlap underlines the importance of SSS-Dent lines for the establishment of the University of Hohenheim's modern IOD-Dent population.

The historical SSS-Dent inbred line B73 was extremely successful as parent of commercial hybrids and is recognized as the dominant source of modern SSS-Dent germplasm in US breeding programs (Mikel and Dudley 2006; Mikel 2008). However, A632 was identified as the most important ancestor of our SSS-Dent germplasm, with B73 having only a comparatively small contribution. Cool climatic conditions during spring and autumn make early maturity a very important breeding goal in Central Europe (Bhosale et al. 2007). This explains the prominent role of A632, which has an early maturity Stiff-Stalk background (Mikel and Dudley 2006). However, this lines is among the important ancestors of modern US SSS-Dent germplasm too (Mikel and Dudley 2006). Other prominent key ancestors of our SSS-Dent germplasm group were: the Canadian line Co125 with unknown genetic background (Messmer et al. 1993), W59E a yellow Dent line from Wisconsin, Oh43, which is a founder of a distinct US Dent sub-group (Mikel and Dudley 2006; Nelson et al. 2008), Wf9, which was released already in 1937 and does not belong to any of the established heterotic groups within Dent (Gethi et al.

2002) and Mo17, a prominent Lancaster line. This wide assembly of key ancestors reflects the diverse and somewhat diffuse genetic background of Central European Dent germplasm. Based on the analysis of pedigrees of lines in the public domain, Rebourg et al. (2003) also assigned an important role to A632 and Co125 in the establishment of the European Dent heterotic group employed in maize breeding programs in Central Europe.

Contributions to modern germplasms and probabilities of gene origin

Our results show that a randomly picked gene in IOD-Dent and SSS-Dent originated with probabilities 90 and 87 %, from one of six and seven key ancestors, respectively. Further, the marginal contribution of the key ancestors varied considerably. This was most pronounced for the IOD-Dent germplasm, where line IOD-0663 contributed almost 50 %. This means that a randomly picked gene in modern IOD-Dent lines originated from this line with 50 % probability.

There are some conceptual differences between F_a on one hand and the effective population size N_e on the other hand. One is that N_e remains constant over generations as long as evolutionary forces, namely the selection intensity, remain constant. The effective number of ancestors F_a, however, is expected to decrease over time under selection, even if the selection intensity remains constant. Nonetheless, both measures are related (Boichard et al. 1997). In particular, the smaller the ratio F_a/n, the smaller N_e (Boichard et al. 1997). Thus, N_e of the University of Hohenheim's IOD-Dent germplasm pool can be expected to be extremely low, owing to the disproportional contribution of IOD-0663.

The low F_a values, especially for the IOD-Dent population, underline the need for broadening the germplasm base by introgression of novel adapted and unadapted genetic resources. These can be chosen mainly to complement the key ancestors. Knowledge about key ancestors can therefore help in identifying new crossing parents in a more systematic manner.

Alternative approaches

We presented a marker-based PGO approach for identifying key ancestors of plant breeding populations. Other approaches have been described in the literature. Mikel and Dudley (2006) used pedigree data to infer the importance of historical US inbred lines with expired plant variety protection act ('ex-PVPA'). However, pedigrees in plant breeding can be vague and often erroneous (Nelson et al. 2008). The unavailability of accurate and complete pedigree information for the breeding program of the University of Hohenheim was precisely the reason why we investigated the use of the marker-based genomic kinship matrix for computing PGO. Another advantage of using genomic data is that deviations from pedigree relationships due to selection can be captured (Lorenz and Hoegemeyer 2013). Several authors therefore used principal coordinate analysis and other marker-based clustering methods for identifying dominant maize lines or reconstructing lineages (Gethi et al. 2002; Liu et al. 2003; Nelson et al. 2008; Lorenz and Hoegemeyer 2013). In our study, the importance of line IOD-0663 for the modern IOD-Dent population of the University of Hohenheim for example, could easily be deduced from the principal coordinates. However, while marker-based clustering can be used to identify key ancestors, it does not provide a quantification of their importance in the way PGO approaches do. Model-based clustering methods, such as implemented in software 'Structure' (Pritchard et al. 2000), can be used to quantify the contribution of distinct sub-populations to current individuals. This method was used by Nelson et al. (2008) for identifying prominent maize inbred lines that represented these sub-populations. The approach of Nelson et al. (2008) is therefore only applicable if the representative founders of these sub-populations are known a priori. Analysis of haplotype sharing was proposed as a method for tracing lineages and identifying essentially derived lines (Romero-Severson et al. 2001) and was used for quantifying ancestry between historical and modern US maize lines, too (van Heerwaarden et al. 2012).

The PGO method proposed by Goddard and Hayes (2009b) presents a straight-forward alternative to these previous approaches, with easily interpretable results. Our proof-of-concept study successfully demonstrated that PGO computed from the genomic kinship matrix can be used for identifying key ancestors in inbred line populations, elucidating the history of breeding programs, and choosing novel germplasm sources and informative genotypes for NGS. Another possible application is to identify key ancestors of specific genomic regions. When the 'key ancestry' of a specific region strongly differs from genome-wide trends or from the set of key ancestors derived by pedigree analysis, this might indicate selection for a particular feature of a historical line. However, given the typically low numbers of effective ancestors and population sizes, local effects of selection might be difficult to distinguish from sampling effects, i.e., drift. In addition to domesticated species, our method might also be suitable for identifying key ancestors of wildlife populations, as long as DNA material of past generations is still accessible. This is certainly the case for, e.g., tree species, which can have enormously long life spans (Lanner 2002).

Author contributions FT conducted the computer simulation and statistical analysis, TAS provided the marker

data, TAS and WS compiled the list of modern and historical inbred lines, FT and AEM devised the study and wrote the manuscript.

Acknowledgments This research was funded by the German Federal Ministry of Education and Research within the Agro-Cluster "Synbreed—Synergistic plant and animal breeding" (grant 0315528D). We acknowledge the contributions of the late Prof. W.G. Pollmer and Dr. D. Klein in the development of the maize inbred lines analyzed in this study and grants from the University of Hohenheim for the development of the modern maize germplasm. We are indebted to Dr. Eva Bauer, Technische Universität München, for generating part of the SNP data used in this study under the "Cornfed" Project, supported by the German Federal Ministry of Education and Research (grant number 0315461). This paper is dedicated to one of the pioneers of biometrical methods in plant breeding, Prof. Dr. H. Friedrich Utz, on the occasion of his 75th anniversary.

Conflict of interest The authors declare no conflict of interest.

References

Bhosale SU, Rymen B, Beemster GTS, Melchinger AE, Reif JC (2007) Chilling tolerance of Central European maize lines and their factorial crosses. Ann Bot 100:1315–1321

Boichard D, Maignel L, Verrier E (1997) The value of using probabilities of gene origin to measure genetic variability in a population. Genet Sel Evol 29:5–23

Cox T, Lookhart G, Walker D, Harrell L, Albers L, Rodgers D (1985) Genetic relationships among hard red winter wheat cultivars as evaluated by pedigree analysis and gliadin polyacrylamide gel electrophoretic patterns. Crop Sci 25:1058–1063

Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams JA, Goddard ME (2011) Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189:317–27

Eding H, Meuwissen THE (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J Anim Breed Genet 118:141–159

El-Kassaby YA, Cappa EP, Liewlaksaneeyanawin C, Klápště J, Lstibrek M (2011) Breeding without breeding: is a complete pedigree necessary for efficient breeding? PLoS One 6: e25737

Elshire RJ, Glaubitz JC, Sun Q, Ja Poland, Kawamoto K, Buckler E, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Addison Wesley Longman Limited, Harlow, p 58

Ganal MW, Durstewitz G, Polley A, Bérard 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 (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(12): e28334

Gerdes JT, Behr CF, Coors JG, Tracy WF (1993) Compilation of North American maize breeding germplasm. CSSA, Madison

Gethi JG, Labate JA, Lamkey KR (2002) SSR variation in important US maize inbred lines. Crop Sci 42:951–957

Goddard ME, Hayes BJ (2009a) Genomic selection based on dense genotypes inferred from sparse genotypes. Proc Assoc Advmt Anim Breed Genet 18:26–29

Goddard ME, Hayes BJ (2009b) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391

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

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

Hayes BJ, Bowman PJ (2011) Accuracy of genotype imputation in sheep breeds. Anim Genet 43:72–80

Kimura M (1964) Diffusion models in population genetics. J Appl Probab 1:177–232

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, Sulem P, Mouy M, Jonsson F, Thorsteinsdottir U, Gudbjartsson DF, Stefansson H, Stefansson K (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 40:1068–1075

Lanner RM (2002) Why do trees live so long? Ageing Res Rev 1:653–671

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

Lorenz A, Hoegemeyer T (2013) The phylogenetic relationships of US maize germplasm. Nat Genet 45:844–845

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

Messmer MM, Melchinger AE, Lee M, Woodman WL, Lee EA, Lamkey KR (1991) Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population: comparison of allozyme and RFLP data. Theor Appl Genet 83:97–107

Messmer MM, Melchinger AE, Boppenmaier J, Brunklaus-Jung E, Herrmann RG (1993) Relationship among early European maize inbreds: I. Genetic diversity among Flint and Dent lines revealed by RFLPs. Crop Sci 32:1301–1309

Mikel MA (2008) Genetic diversity and improvement of contemporary proprietary North American dent corn. Crop Sci 48:1686–1695

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

Montana G (2005) HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. Bioinformatics 21:4309–4311

Nelson PT, Coles ND, Holland JB, Bubeck DM, Smith S, Goodman MM (2008) Molecular characterization of maize inbreds with expired US plant variety protection. Crop Sci 48:1673–1685

Pausch H, Aigner B, Emmerling R, Edel C, Götz KU, Fries R (2013) Imputation of high-density genotypes in the Fleckvieh cattle population. Genet Sel Evol 45:3. doi:10.1186/1297-9686-45-3

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

R Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/, ISBN 3-900051-07-0

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

Romero-Severson J, Smith JSC, Ziegle J, Hauser J, Joe L, Hookstra G (2001) Pedigree analysis and haplotype sharing within diverse groups of Zea mays L. inbreds. Theor Appl Genet 103:567–574

Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegle J (1997) An evaluation of the utility

of SSR loci as molecular markers in maize (*Zea mays L.*): comparisons with data from RFLPS and pedigree. Theor Appl Genet 95:163–173

Stich B, Melchinger AE, Piepho HP, Hamrit S, Schipprack W, Maurer HP, Reif JC (2007) Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. Theor Appl Genet 115:529–536

Technow F (2012) hypred: simulation of genomic data in applied genetics. R package version 2

Technow F, Bürger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 3:197–203

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197:1343–1355

van Heerwaarden J, Hufford MB, Ross-Ibarra J (2012) Historical genomics of North American maize. PNAS 109:12420–12425

Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120:1289–1299

Wright S (1978) Evolution and genetics of populations, vol IV. The University of Chicago Press, Chicago