Theor Appl Genet (2014) 127:559–571
DOI 10.1007/s00122-013-2239-x

ORIGINAL PAPER

Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions

Ibraheem Adetunji · Glenda Willems · Hendrik Tschoep · Alexandra Bürkholz · Steve Barnes · Martin Boer · Marcos Malosetti · Stefaan Horemans · Fred van Eeuwijk

Received: 21 May 2013 / Accepted: 20 November 2013 / Published online: 1 December 2013

Abstract
Key message **Linkage disequilibrium decay in sugar beet is strongly affected by the breeding history, and varies extensively between and along chromosomes, allowing identification of known and unknown signatures of selection.**
Abstract Genetic diversity and linkage disequilibrium (LD) patterns were investigated in 233 elite sugar beet breeding lines and 91 wild beet accessions, using 454 single nucleotide polymorphisms (SNPs) and 418 SNPs, respectively. Principal coordinate analysis suggested the existence of three groups of germplasm, corresponding to the wild beets, the seed parent and the pollen parent breeding pool. LD was investigated in each of these groups, with and without correction for genetic relatedness. Without correction for genetic relatedness, in the pollen as well as the seed parent pool, LD persisted beyond 50 centiMorgan (cM) on four (2, 3, 4 and 5) and three chromosomes (2, 4 and 6), respectively; after correction for genetic relatedness, LD decayed after <6 cM on all chromosomes in both pools. In the wild beet accessions, there was a strong LD decay: on average LD disappeared after 1 cM when LD was calculated with a correction for genetic relatedness. Persistence of LD was not only observed between distant SNPs on the same chromosome, but also between SNPs on different chromosomes. Regions on chromosomes 3 and 4 that harbor disease resistance and monogermy loci showed strong genetic differentiation between the pollen and seed parent pools. Other regions, on chromosomes 8 and 9, for which no a priori information was available with respect to their contribution to the phenotype, still contributed to clustering of lines in the elite breeding material.

Communicated by J. Yan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2239-x) contains supplementary material, which is available to authorized users.

I. Adetunji · G. Willems · H. Tschoep (✉) · A. Bürkholz · S. Barnes · S. Horemans
SESVanderHave, Soldatenplein 15, 3300 Tienen, Belgium
e-mail: hendrik.tschoep@sesvanderhave.com

M. Boer · M. Malosetti · F. van Eeuwijk
Biometris, Applied Statistics, Wageningen University, PO BOX 100, 6700 AC Wageningen, The Netherlands

Introduction

The beet plant (*Beta vulgaris* L.) originates from the Mediterranean regions and has been cultivated mainly for the nutritional qualities of its leaves (Cooke and Scott 1993; Panella and Lewellen 2007). *Beta vulgaris* contains many subspecies such as leaf beets, fodder beets, red beets, sea beets, sugar beets and Swiss chards. Sugar beet (*B. vulgaris* ssp. *vulgaris*) is a biennial plant and is widely grown in temperate regions (Draycott 2006). It accounts for approximately 25 % of the world's sugar production (Draycott 2006). The domestication of sugar beet started late, in the eighteenth century, and is, therefore, relatively recent when compared to that of other major crop plants (Cooke and Scott 1993). Commercial sugar beet varieties are mainly developed as diploid hybrids by creating an F1 between two genetically diverse inbred lines (seed and pollen parent lines) with a yield superior to both parents.

This phenomenon is commonly known as hybrid vigor or heterosis (Shull 1908; Falconer and Mackay 1996). Introgression of important characteristics such as monogermy, maintainers of cytoplasmic male sterility (CMS) or resistance to the *Beet necrotic yellow vein virus* (BNYVV)

into elite breeding lines can be considered as genetic bottlenecks that are likely to have shaped the linkage disequilibrium (LD) pattern and, therefore, the population structure of the sugar beet breeding material.

The extent of LD has been investigated in elite sugar beet lines (Kraft et al. 2000; Li et al. 2010, 2011) as well as in wild beet accessions (Desplanque et al. 2000; Arnaud et al. 2003, 2009; Fievet et al. 2007). Kraft et al. (2000) reported significant LD between markers that were <3 centiMorgan (cM) apart using nine sugar beet elite lines and a set of 451 AFLP markers. In a germplasm set consisting of approximately 200 inbred lines belonging to the seed and pollen parent heterotic pool, Li et al. (2010) reported LD decay to an r^2 value of 0.1 or less within 10 cM using 23 SSR markers. Moreover, they also observed significant LD between loci on different chromosomes, indicating that LD in sugar beet can be generated by factors such as population structure, genetic drift and familial relatedness, as stated by Li et al. (2010). A recent study of LD in an elite sugar beet germplasm set composed of 264 yield- and 238 sugar-type pollen parent inbred lines reported significant LD between loci pairs that were 7 cM apart (Li et al. 2011). In the groups composed of either yield- or sugar-type lines; however, LD was much more extensive and still significant for loci at a distance of 45 and 21 cM, respectively. The variation in LD for the two groups was attributed to differences in selection history.

Studies looking at the extent of LD in sugar beet showed good prospects for genome-wide association mapping studies (GWAS) (Stich et al. 2008a, b; Würschum et al. 2011a, b). Simultaneously, studies by Li et al. (2010, 2011) showed that corrections for differential relatedness, either using kinship or groupings, are required to control the number of false positive marker-trait associations detected in GWAS. These corrections were extensively discussed for *Arabidopsis thaliana* by Zhao et al. (2007) and Atwell et al. (2010). Various statistical approaches have been developed to assess population structure in crop species, as well as natural populations. The most widely used is the model-based clustering program STRUCTURE (Pritchard et al. 2000), whose goal is to identify groups and assign individuals to these different groups using an estimated membership coefficient. Other, distance-based, approaches include principal coordinate analysis (PCoA), and principal component analysis (PCA) (Patterson et al. 2006).

Several reports on LD decay, and genetic diversity have been published on elite sugar beet germplasm (Jung et al. 1993; Kraft et al. 1997, 2000; McGrath et al. 1999; Smulders et al. 2010; Li et al. 2010, 2011), though using different marker types, highly variable marker numbers, and germplasm sets. To our knowledge, none of these studies provided a direct comparison of LD patterns in cultivated and wild beets using the same or similar marker sets. In this study, we analyzed population structure in a combined set of wild beet accessions and elite sugar beet lines genotyped at 459 markers. In the prospect of GWAS, we investigated the patterns of LD decay on a local and on genome-wide basis on a set of 233 elite sugar beet breeding lines belonging to different heterotic groups, and a set of 91 wild *B. vulgaris* accessions. We also investigated the effect of differential genetic relatedness on LD decay in wild and cultivated beets, when estimating LD between SNPs. To our knowledge, unlike in other crop species, selective sweep analyses (Clark et al. 2004; Palalsa et al. 2004; Olsen et al. 2006) are currently lacking in sugar beet. In this study, to quantify the genetic differentiation between the pollen and seed parent group and to identify possible signatures of selection, we calculated F_{ST} values (Weir and Cockerham 1984). Moreover, using a Z-test for comparing proportions, we tested whether the allele frequencies at marker loci differed significantly between pollen and seed parent pools.

Materials and methods

Plant materials and genetic markers

The germplasm used within this study comprises two major groups: 234 elite sugar beet breeding lines, which are property of SESVanderHave. Additionally, a set of 99 wild beet accessions which were obtained from public databases such as the International Data Base for Beet (IDBB) or National Genetic Resources Program (NGRP) was included in the study. The set included 20 sea beet accessions (*B. vulgaris* ssp. *maritima*) and 79 cultivated beet accessions such as garden, leaf, red or fodder beet (*B. vulgaris* ssp. *vulgaris*) and was composed to cover a wide range of cultivated beet types and sea beets as well as a broad range of geographical origins within the subspecies. The reported sites of collection ranged from the Northern and Western European countries such as Denmark, France, the Netherlands, UK, and Germany, to Mediterranean countries such as Italy, Spain, Turkey, Tunisia and Greece, and to Central and Eastern Europe with collection sites in Poland, Russia, Georgia and Kazakhstan.

The elite sugar beet breeding pool is composed of 139 lines that are used as pollen parent and 95 lines that are used as seed parent. The pollen and seed parents reflect the two different heterotic groups. A total of 498 SNPs identified as polymorphic in a representative group of the SESVanderHave germplasm (composed of 410 sugar beet genotypes representing historic and currently used sugar beet lines used as pollinators or as seed parents in breeding schemes) were selected to give a uniform coverage of the genome, as far as possible. The SNPs were designed in

both genomic and expressed sequences (cDNAs) and had previously been mapped using three different F2 mapping populations to obtain the genetic position of the SNPs. The three different mapping populations were established by crossing (1) two diverse lines derived from sugar beet breeding populations ($n = 160$), (2) a French maritima accession and a sugar beet line ($n = 130$), and (3) a sugar beet line and a Swiss chard accession ($n = 160$).

Genotyping was performed on DNA extracted from freeze-dried leaf tissue pooled from four to eight plants for each of the 234 elite sugar beet breeding lines and 99 wild beet accessions in the germplasm collection, using KASPar assays (KBiosciences) for individual SNPs.

For the pools of pollen parents, seed parents and wild beets, genetic diversity was characterized by gene diversity (I), or expected heterozygosity, which is defined as the probability that two randomly selected alleles from a population are different. Furthermore, observed heterozygosity (H) and the polymorphism information content (PIC) were evaluated for each of the three populations. These parameters were calculated with PowerMarker 3.25 (Liu and Muse 2005).

Population structure analysis

Population structure was investigated using two different methods: PCoA and STRUCTURE. PCoA was carried out on a similarity matrix to produce principal coordinate scores which were then used to investigate population subgroups in the germplasm collection. For the calculation of the similarity or kinship matrix, SNPs with a minor allelic frequency (MAF) <0.05 were excluded. This matrix was obtained using the Dice distance as proposed by Nei and Li (1979), which defines similarity between two individuals as the number of common alleles divided by the average number of alleles observed in both individuals. Its values range between 0, when the individuals do not share a single allele across the set of marker loci, and 1, if the individuals are genetically identical at all marker loci used in the analysis. To calculate the similarity matrix, a variant allele and a reference allele were identified for each marker. The coding pattern for each of the alleles is a vector that can take the values 0, 1, or 2, for which 0 designates absence of the variant alleles, 1 designates presence of one of the variant alleles and, 2 designates presence of two variant alleles at the marker locus. These analyses were performed using QKINSHIPMATRIX and PCO and DMST procedures in Genstat 14th edition (VSN International 2011). The software STRUCTURE (Pritchard et al. 2000) was run, assuming a population admixture model with independent allele frequencies between subgroups. Ten replications were run for each of the subpopulation numbers, K, that were chosen to range from 1 to 9. Each run had 200,000 MCMC iterations, of which the first 100,000 iterations (that were used to monitor whether a chain reached stationarity) were discarded as burn in. The delta K method was used to identify the number of subgroups in the dataset (Evanno et al. 2005). Differences between identified groups were tested by analysis of molecular variance (AMOVA), as implemented in Arlequin 3.5 (Excoffier and Lischer 2010).

Linkage disequilibrium analysis

To establish LD between markers, linear regression models were used with a first SNP as response and a second as predictor, with or without corrections for genetic relatedness (GR). The GR was modeled using the idea of Patterson et al. (2006), where eigenvalue decomposition is performed on the normalized genotype by marker score matrix. All significant principal components (1…Q) are used to correct for GR by introducing them as covariates in the linear model of one marker on another. The linear regression models, without and with correction for GR, are thus given by

$$\begin{cases} Y_i = \mu + \beta X_{im} + \varepsilon_i \\ Y_i = \mu + \sum_{q}^{Q} \alpha_q C_{iq} + \beta X_{im} + \varepsilon_i \end{cases},$$

where Y_i and X_{im} represent the marker scores for individual i for the response and the mth predictor SNP, respectively. C_{iq} contains the principal component scores for individual i and component q, β and α_q are the regression coefficients for the regressor SNP and the principal components, respectively. Markers were investigated for their LD up to a distance of 50 cM. ε_i stands for the error terms, which are assumed to be normally distributed $N(0, \sigma^2)$. The estimate of the LD, r^2, was obtained from fitting the above models and calculating ratios of sums of squares for explained variation by the mth SNP to total sum of squares, without and with correction for GR (Mangin et al. 2012):

$$\begin{cases} r^2 = \frac{\mathrm{SSR}(X_m)}{\mathrm{SST}} \\ r^2 = \frac{\mathrm{SSR}(X_m|C_1 \ldots C_m)}{\mathrm{SST}} \end{cases},$$

where $SSR(X_m|C_1 \ldots C_m)$ is the sum of squares explained by marker m given that the significant principal component scores are already in the model, $\mathrm{SSR}(X_m)$ is the sum of squares for marker m without correction for GR and SST is the total sum of squares. For the estimation of LD decay, LD was calculated between SNPs at maximally 50 cM. A non-linear quantile regression was fitted to the response variable, r^2 values. The 95th percentile curve from the resulting non-linear quantile regression was plotted alongside with the r^2 values against the genetic distance x:

$$r^2 = Ae^{-\frac{x}{C}} + B,$$

where A and B are linear constants, and C is a non-linear decay constant. The LD decay was obtained by solving this equation for a given threshold for r^2, denoted by r^2_{Thr}:

$$r^2_{\mathrm{Thr}} = Ae^{-\frac{x}{C}} + B \rightarrow x = \frac{-\ln((r^2_{\mathrm{Thr}} - B)/A)}{C}$$

The r^2_{Thr} was obtained by random sampling of 10,000 pairs of unlinked markers, calculating the r^2 with correction for GR, and taking the 0.999 quantile as threshold.

Furthermore, for investigation of local LD patterns, the median LD was calculated for each SNP with all SNPs within a window of 20 cM on either side, without correction for GR.

Identification of selective sweeps

Two different approaches, a genome-wide calculation of F_{ST}'s and a test for comparison of allele frequencies using Z tests, were used to identify chromosomal regions showing differentiation and, therefore, likely under selection in the seed or in the pollen parent pools. Wright's F_{ST} was calculated according to Weir and Cockerham (1984) with the equation given as

$$F_{\mathrm{ST}} = 1 - \frac{H_{\mathrm{s}}}{H_{\mathrm{T}}}$$

where H_{T} corresponds to the heterozygosity expected under Hardy–Weinberg equilibrium (HWE) without subdivision, and H_{S} corresponds to the expected average heterozygosity of the two groups assuming HWE.

To identify SNPs showing a significant deviation between allele frequencies in the pollen and the seed parent group, we calculated the test statistic Z which is given by

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},\ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2},$$

where $p_1$and $p_2$are the sample allele frequencies for the two populations, x_1 and x_2 are the number of times the most frequent allele appears in each sample, and n_1 and n_2 are the sample sizes (Murray and Larry 2007).

A family-wise error rate was used to correct for multiple testing, using a modified Bonferroni approach. The significance level ($\alpha = 0.05$) was divided by the number of independent markers across the genome, estimated as $n = 286$, defined as the sum of independent markers across chromosomes. The number of independent markers per chromosome was calculated by dividing the length of the chromosome by the estimate for LD decay obtained in this study. Any SNP with a test statistic Z greater than the critical value of $|Z_{\alpha/2n}| = 3.75$ was considered to have different frequencies in seed and pollen elite lines.

Results

Summary of SNPs

The initial dataset consisted of 333 individual entries which contained 234 elite sugar beet lines and 99 wild beet accessions that were genotyped with 498 SNPs. After excluding markers that have (1) more than 25 % missing values across all genotypes, (2) a MAF $\leq$0.05 and (3) genotypes that have more than 40 % missing values over all the markers, the whole dataset was reduced to 324 lines corresponding to 233 elite sugar beet lines and 91 wild beet accessions genotyped with 459 SNPs. These markers represent 203 different mapping positions and their distribution

Table 1 Descriptive statistics of SNP markers: chromosome, length of chromosome expressed in centiMorgan (cM), SNP, the number of markers per chromosome after removing markers with minor allelic frequency (MAF) <0.05 and with >40 % missing values, number of loci with MAF <5 %, number of loci with MAF between 5 and 10 %, 95th percentile gives the distance between markers below which 95 % of the distances occurs, Loci >1 SNP gives the number of loci which have two or more SNPs at the same genetic map position, Loci = 1 SNP gives the number of loci which have only one SNP at a position

Chrom	Length[a]	SNP	MAF < 5 %	5 % < MAF < 10 %	95th distance percentile	Loci > 1 SNP	Loci = 1 SNP
c1	66	41	2	2	5.5	12	11
c2	61	50	4	3	6.8	12	3
c3	89	70	5	1	6.0	21	7
c4	83	73	5	3	5.0	20	8
c5	61	54	2	1	4.6	17	8
c6	79	43	0	3	7.6	12	9
c7	59	32	1	2	8.7	11	6
c8	77	44	1	2	8.2	15	9
c9	51	52	0	4	5.7	19	3
		459				139	64

over all nine individual chromosomes is shown in Table 1. Applying the above criteria separately to the elite sugar beet breeding lines and the wild beet accessions datasets, the elite sugar beet breeding lines dataset was reduced to 454 SNPs that represent 202 different mapping positions, while the wild beet accessions dataset reduced to 418 SNPs that represent 190 different mapping positions. Splitting the elite pool into heterotic groups gave 95 seed and 138 pollen parent breeding lines that were genotyped with the same set of 459 SNPs. On average, the gene diversity index and PIC statistics were larger in the wild accessions than in the elite material, while the difference between pollen parent and seed parent breeding lines was small (Suppl. Table 1). This implies more genetic variability within the wild accessions group than within the elite material. The level of heterozygosity was more than twofold higher in the wild beet accessions than in the elite breeding lines (Suppl. Table 1). The pollen and seed parents showed a comparable level of heterozygosity.

Identification of population subgroups

To investigate population structure, PCoA was performed on the entire dataset of 324 genotypes. The PCoA based on the similarity matrix explained 14.44 and 6.98 % of the genetic variation with the first and the second PCoA axis, respectively. Plotting the scores of the two first PCoA axes confirmed the presence of three groups within the germplasm used in this study, though with some overlap between pollen parent lines, seed parent lines and wild beets (Fig. 1). On the first axis, the majority of the pollen parent lines are separated from the other lines. On the second axis, most elite lines belonging to the seed parent heterotic pool cluster distinctly from wild beet accessions, while the remaining seed parent lines and also few pollen parent breeding lines are clustered together with the wild beet accessions. The wild beet accessions are displayed as a single group, since PCoA did not reveal a clear differentiation between wild accessions previously described as *B. vulgaris* ssp. *vulgaris* and ssp. *maritima* (data not shown). Some overlap existed between pollen parent lines, seed parent lines and wild beets (Fig. 1). Plotting higher PCoA axes did not reveal an improved separation of the known groups (data not shown). As a second analysis on differentiation, we used STRUCTURE and observed a gradual increase in log likelihood from $k = 1$–9 (Suppl. Fig. 1A). The largest delta k was detected at $k = 3$ which suggests the presence of three distinct groups (Suppl. Fig. 1B; Suppl. Fig. 2). The AMOVA result showed a significant differentiation among these subgroups with approximately 34 % of the total genetic variation explained by the differences between the subgroups (Suppl. Table 2).

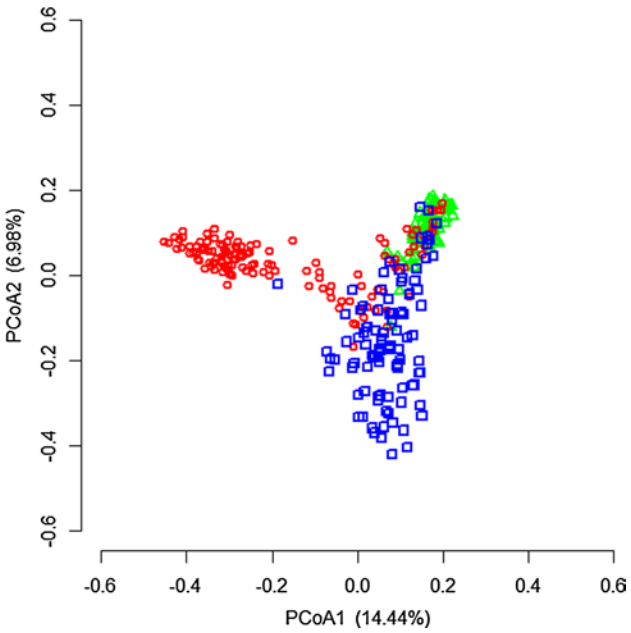

Fig. 1 Plot of the first two axes from a principal coordinates analysis for the elite and wild pools using the 459 SNPs. The *red circles* indicate lines classified as pollen donors, the *blue squares* indicate lines classified as seed parents and the *green triangles* correspond to wild accessions

As the wild beet accessions used within this study include individuals from the ssp. *maritima* as well as *vulgaris,* and since they originate from regions across the whole of Europe, we investigated whether the wild beets belonging to the same subspecies (*maritima* or *vulgaris*) or type (fodder, leaf, garden or red beet) clustered together using the PCoA plot. Interestingly, we could not detect a separation of the *B. vulgaris* ssp. *vulgaris* from the *maritima* subspecies (data not shown). Moreover, no clusters were found within the *B. vulgaris* ssp. *vulgaris* that coincided with database annotations as leaf, fodder, red or garden beet (data not shown). Finally, PCoA using exclusively wild beet accessions also failed to identify clusters representing the sites of collection (data not shown). These results were rather surprising and might be due to (1) too small sample sizes, (2) the incorrect annotation of the real site of origin or type of wild beet in the databases and (3) a bias in the SNPs, as these were selected on a set of elite breeding lines.

Analysis of linkage disequilibrium

LD decay was estimated in the pollen parent pool, the seed parent pool and wild accessions, independently. The threshold, r^2_{Thr}, based on sampling 10,000 pairs of unlinked markers, was 0.17 for the pollen parent pool, and 0.18 for the seed parent pool, while the wild accessions pool had a value of

Table 2 LD decay distance in cM for pollen parent pool, seed parent pool, and wild accessions

Chromosome	LD decay distance[a]					
	Without population structure correction			With population structure correction		
	Pollen parent	Seed parent	Non elite	Pollen parent	Seed parent	Non elite
1	9.5	6.9	6.3	5.3	3.4	3.7
2	>50	>50	0.7	4.9	2.3	0.1
3	>50	6.1	1.8	4.7	3.1	1.6
4	>50	>50	1.0	2.3	3.5	0.3
5	>50	4.1	0.1	5.4	2.9	0.3
6	9.6	>50	1.2	5.8	3.5	0.4
7	28.7	6.8	0.1	4.5	4.5	0.1
8	7.7	6.3	1.9	2.6	1.9	1.2
9	4.0	3.2	2.0	1.9	2.5	1.8

[a] LD decay distance expressed in cM

0.15. When GR was not accounted for, LD extended beyond 50 cM on four chromosomes in the pollen parent pool, and three chromosomes in the pollen parent pool (Table 2). When GR was corrected for by including the significant principal component scores (3 and 2 axes, respectively) for pollen and seed parent pools in the model, LD decreased faster with increasing genetic map distance on all chromosomes within both heterotic groups. In the pollen and seed parent pools, LD decayed within a distance of <6 and 4 cM on all chromosomes, respectively (Table 2). As shown in Fig. 2 for chromosome 3, LD decay distance was much greater in both heterotic groups than in the wild beet accessions and this was true for all other chromosomes (Table 2) with and without correction for GR. Interestingly, correction for GR made almost no difference in LD decay within the wild beet accessions (Table 2). This is surprising, but consistent with the observation that there were neither clusters according to geographical origin nor to the phenotypic type description (e.g. leaf beet, garden beet, red beet and fodder beet) in the wild beet accessions. It suggests that either there is no clear genetic structure within the wild beet accessions, or that the sample size may have been too small to pick up any population structure with the SNPs used in this study.

With the aim of studying local LD patterns in the pollen and seed parent heterotic pool, we calculated median r^2 values at every SNP using the r^2 values with markers within a window of 20 cM. Most pronounced median r^2 estimates were found on chromosomes 1 and 3 for the pollen parent pool and on chromosome 4 and 9 in the seed parent heterotic group (Fig. 3). The maximum value of the median r^2 estimates was found at the top of chromosome 9 in the seed parent pool. Some of the regions on chromosomes 3 and 4 showing high r^2 estimates coincide with regions that are known to be under strong selection in elite sugar beet germplasm such as the loci harboring the maintainers of CMS and the *Rz1* locus conferring resistance to BNYVV (Owen 1945; Barzen et al. 1992; Pillen et al. 1993; Scholten et al. 1999; Schondelmaier and Jung 1997; Lein et al. 2007; Hagihara et al. 2005). To verify the extent of long-range LD within the two sugar beet heterotic pools, we calculated LD between all possible pairs of loci and displayed this in a heatmap (Fig. 4a, b). In the pollen parent pool, we observed r^2 values above 0.2 (below 0.4) between SNPs located on chromosome 3 and SNPs on six out of the remaining eight chromosomes (Fig. 4a). It is interesting to note that this chromosome harbors loci involved in resistance to rhizomania (Barzen and Mechelke 1995; Scholten et al. 1999; Grimmer et al. 2007), a trait that has been heavily selected in the pollen parent heterotic group. This was not seen in the seed parent pool where only in a few cases LD between SNPs located on different chromosomes reached the levels observed in the pollen parent pool (Fig. 4b). Figure 4a, b shows that the strongest LD was detected between SNPs located close to each other on the same chromosome. Only in the pollen parent pool on chromosome 3, r^2 values close to or higher than 0.6 were also observed between SNPs located further away (Fig. 3).

Identification of selective sweeps

A genome-wide calculation of F_{ST} was performed to assess genetic divergence between the pollen and seed parent heterotic pools. Regions that are genetically divergent between the breeding pools, such as the CMS maintainer loci and the *Rz1* locus which are only present in the seed and pollen parent plants, respectively, might indicate regions under selection in one or both of them. SNPs showing high F_{ST} values were detected on all nine chromosomes of the sugar beet genome (Fig. 5). Twenty-four SNPs on chromosomes 3 (14 SNPs), 4 (6 SNPs), 8 (1 SNP) and 9 (3 SNPs) showed an F_{ST} value above 0.25. The highest F_{ST} value (0.54) was obtained at a SNP located on chromosome 4 at 82 cM. Allele frequencies at all SNPs in both heterotic pools were compared by *Z* tests in an additional analysis for indications

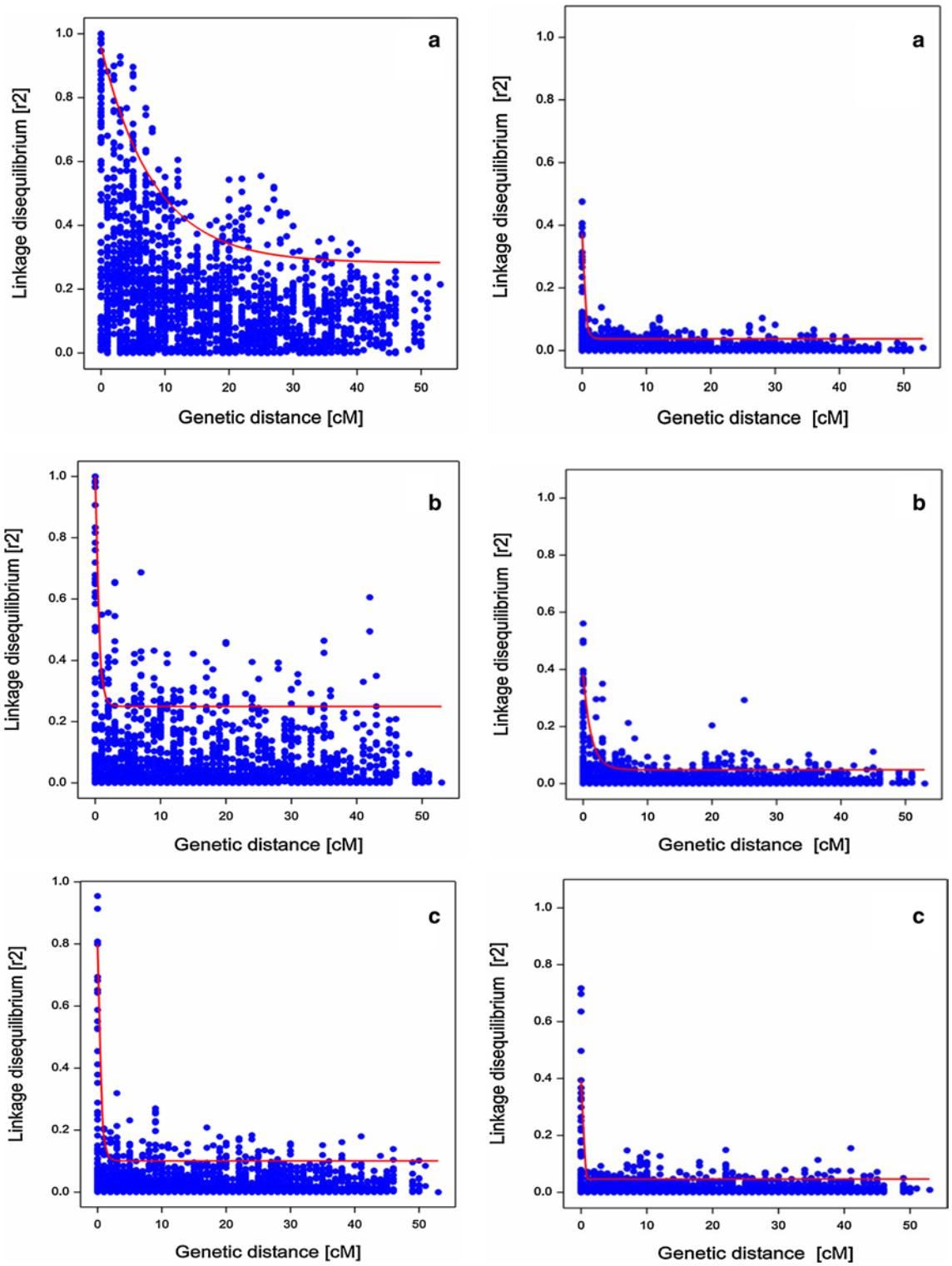

Fig. 2 Linkage disequilibrium (r^2) between SNP markers as a function of genetic map distance for chromosome 3 for pollen parent (**a**), seed parent (**b**) and the wild beet accessions (**c**). *Left panel* plot of r^2 as a function of genetic map distance (cM) without correcting for genetic relatedness. *Right panel* plot of r^2 as a function of genetic map distance (cM) with correction for genetic relatedness using principal component scores. The *red curve* corresponds to the 95th percentile of r^2 estimates between SNP markers

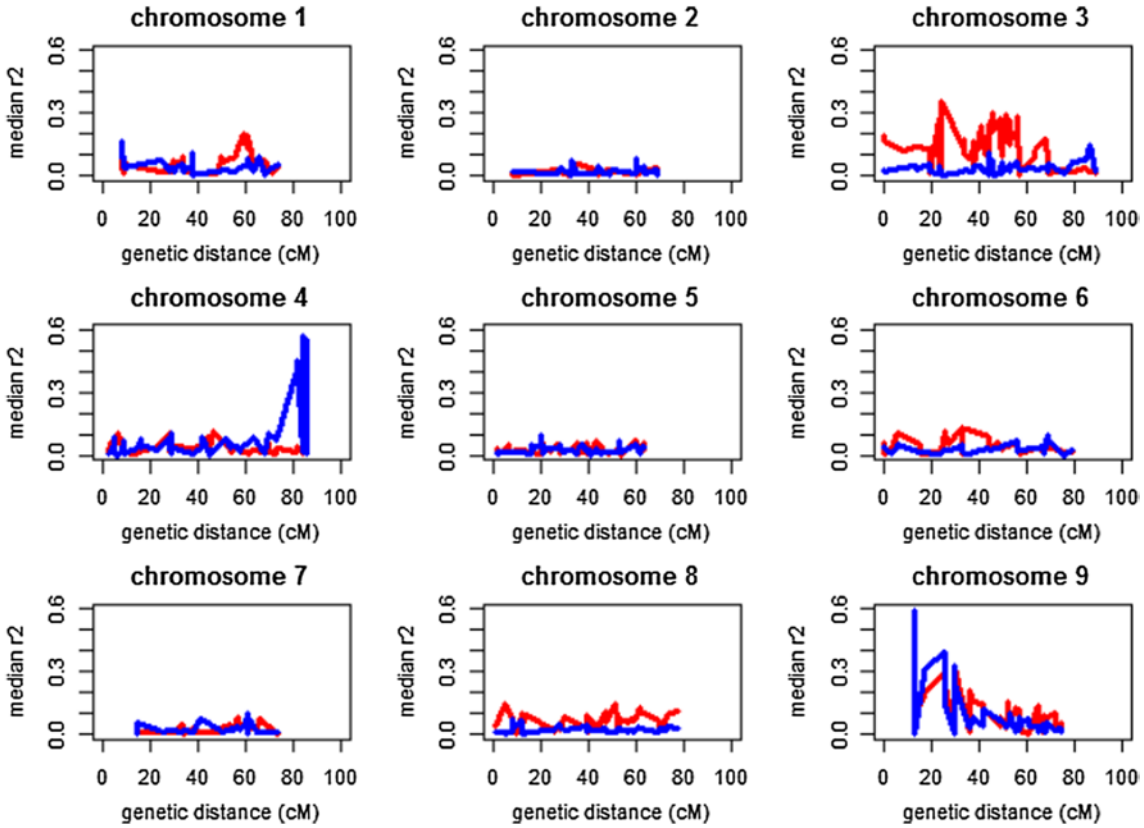

Fig. 3 Median r^2 estimates between markers and their direct neighbors (within 20 cM) along the chromosomes with no correction for genetic relatedness for the pollen parent pool (*red line*), and the seed parent pool (*blue line*)

of signatures of selection in the pollen and/or seed parent pool. Significant differences in allele frequencies between both pools were found on all nine chromosomes. The 24 SNPs identified previously with the F_{ST} approach were also among the significant SNPs identified with the *Z*-test. Chromosome 3 showed the highest number of SNPs with significant allele frequency differences between pollen parent and the seed parent pool (50 SNPs), followed by chromosomes 4 (40 SNPs) and 9 (33 SNPs). On chromosome 3 with the exception of 10 SNPs, all significant SNPs mapped within 5 cM of each other at the middle of the chromosome. SNPs showing significant allele frequency differences between the seed parent and pollen parent pools were observed at both ends of chromosome 4. Frequency differences on chromosome 9 were lower than those on chromosomes 3 and 4. However, in contrast to chromosomes 3 and 4, SNPs with differential allele frequencies between groups were not concentrated in one region, but scattered over the full length of the chromosome.

To check whether markers showing evidence of selection were sufficient to reveal the clustering pattern observed (Fig. 1), we performed a PCoA using all SNPs that showed a significant allele frequency difference between the elite breeding lines in both heterotic pools. Knowing that certain regions on chromosome 3 as well as some on chromosome 4 harbor genes involved in traits that have been heavily selected in one of the heterotic pools, we excluded the SNPs on these regions to avoid a bias in the analysis. On chromosome 3, loci involved in rhizomania resistance (Barzen et al. 1992; Scholten et al. 1999; Lein et al. 2007), and in maintenance of CMS have been mapped (Schondelmaier and Jung 1997), while on chromosome 4 monogermy (Barzen et al. 1992) and loci contributing to the expression of CMS sterility (Hagihara et al. 2005) are located. When using the remaining 142 SNP markers, for which we have no additional information with regard to the phenotype they are possibly involved in, two clusters were obtained within a PCoA that correspond to the two heterotic pools present in the collection of elite breeding lines (Fig. 6). As expected, PCoA using the SNPs that were not significant in the test for comparison of allele frequencies between the pools did not reveal any clustering of the elite breeding lines (Suppl. Fig. 3).

Discussion

In this work, we have analyzed population structure and genetic diversity in a set of elite sugar beet lines and wild

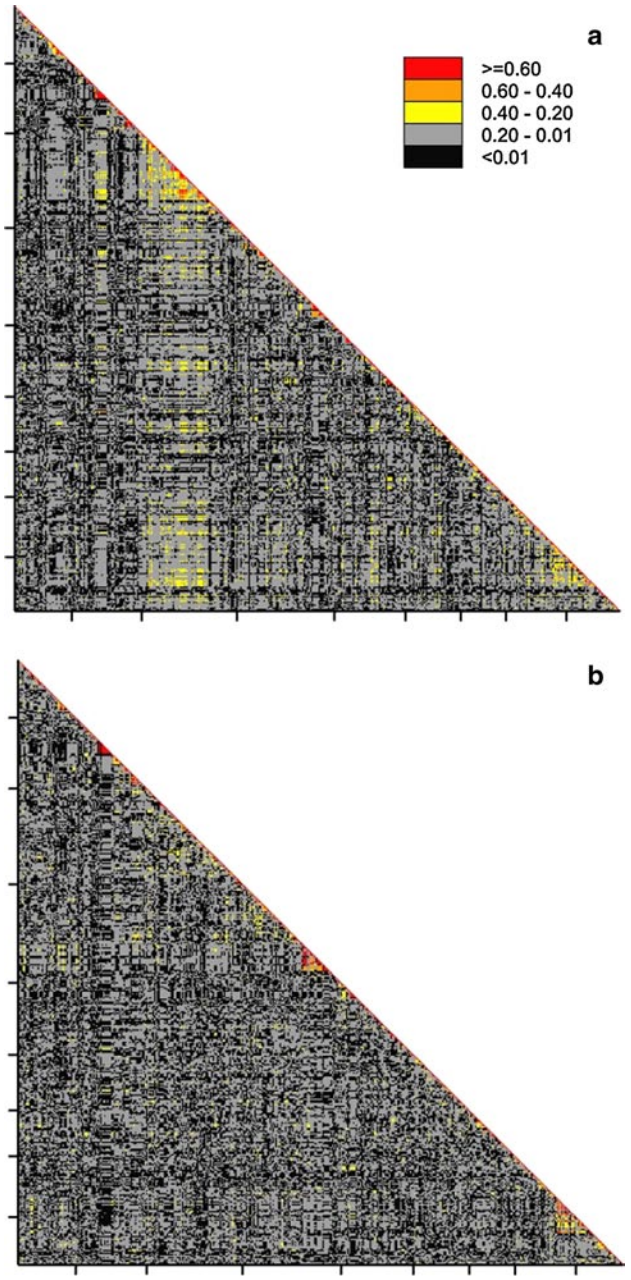

Fig. 4 Heatmap of r^2 values between all possible SNP pairs without correction for genetic relatedness, in the pollen parent pool (**a**) and the seed parent pool (**b**)

beet accessions. Using two different approaches, PCoA and STRUCTURE, we could clearly distinguish three groups in the set of genotypes used in the study. Using a priori information, we were able to verify that these groups corresponded to the two heterotic pools included in the elite germplasm and the wild beet accessions, respectively. However, the PCoA plot also suggested that some elite lines belonging to one heterotic group were finally more closely related to lines belonging to the other heterotic pool than to lines belonging to the same heterotic pool. This is the case for instance for one line classified as seed parent, though it clusters with the pollen donor lines. However, this result is not surprising as this line originated from a pollen donor line. While some pollen parent lines were clearly separated from the majority of the seed parent lines, other elite lines belonging to both heterotic groups clustered rather towards the center of the PCoA plot and even overlapped. These elite lines originated from the opposite heterotic pool and have probably been used in breeding for a shorter time period than the breeding lines that are positioned at the extremes of the PCoA axes, which are supported by the relative short breeding history and long generation time of sugar beet compared to other major crops such as maize. In sugar beet, recurrent selection within the pollen parent heterotic pool is typically faster than in the seed parent pool, leading, on average, to a higher number of recombination events per unit time. The more pronounced separation of the pollen parent lines from the wild beets than the seed parent lines is consistent with this. The clustering of the wild beet accession together with a few breeding lines from the pollen and seed parent heterotic pool can be explained by the fact that part of these lines was recently developed from wild beets.

Without correction for GR, LD between markers within 50 cM of each other remained consistently high over the length of the chromosome in four out of nine chromosomes in the pollen parent pool and three out of nine chromosomes in the seed parent pool. Previous reports on LD decay did not report such extreme results. Kraft et al. (2000) reported strong LD only for markers that were tightly linked, while Li et al. (2010) reported a decay of r^2 to 0.1 within 10 cM on a genome-wide scale. Multiple reasons can be put forward to explain the discrepancies between the observations made in the current and previous studies. The size of the germplasm set, the number of markers, as well as the type of markers used to analyze LD decay were very different between this work and the previous studies, which precludes any direct comparison of LD measures. Finally, we prefer to report LD decay values on a per-chromosome basis rather than on a genome-wide scale, because we know that all chromosomes have not been subjected to the same selection intensity throughout their breeding history. Therefore, we suspected to identify quite contrasting patterns in LD over the different chromosomes, as reported in crops like maize (Yan et al. 2009). This is also what we have found, as no decay in LD at all was observed on four chromosomes in the pollen parent pool and three chromosomes in the seed parent pool. By comparing the LD decay in the heterotic pools with the LD decay, we observed in the set of wild beet accessions, we have confirmed that the LD decay patterns we observed in the elite lines do not reflect the wild origins of sugar beet.

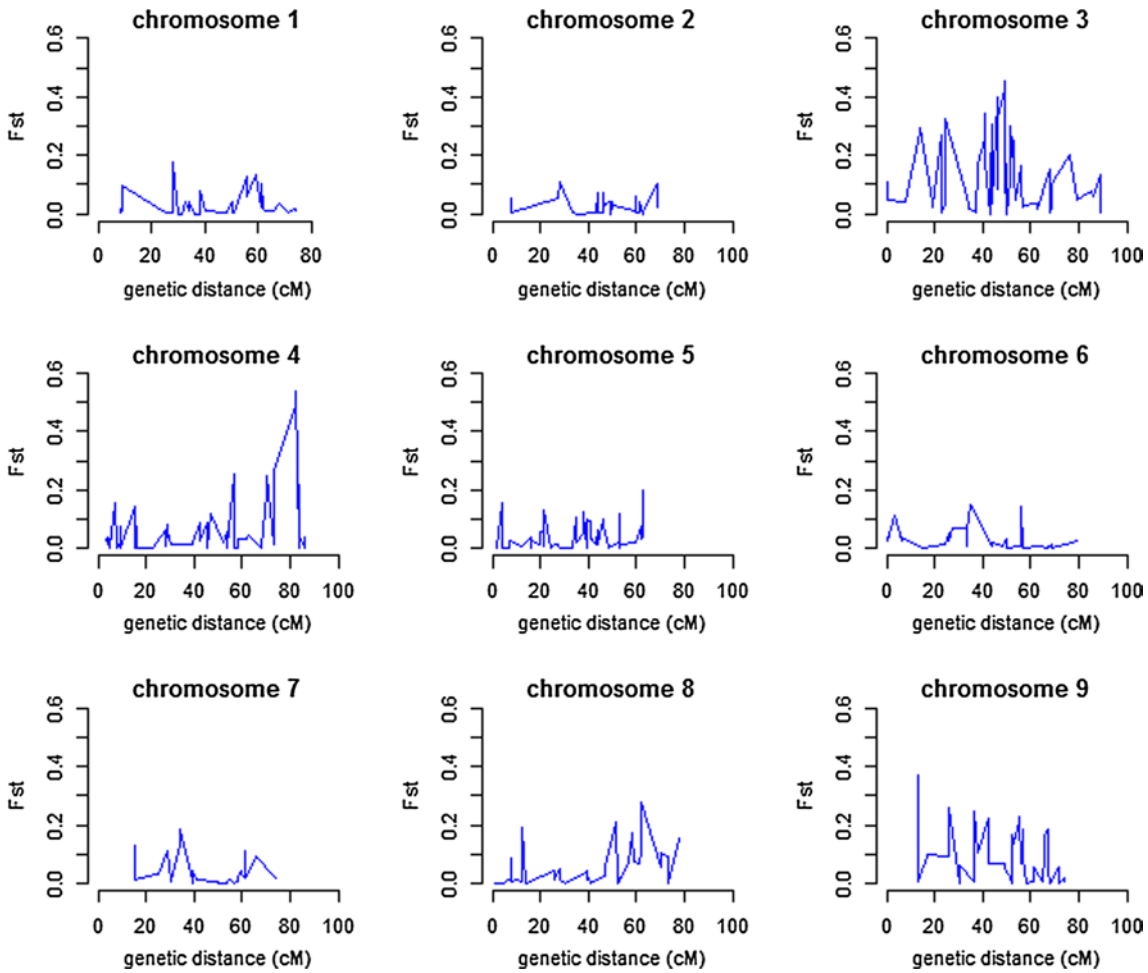

Fig. 5 F_{ST} against genome position for comparison of pollen parent and seed parent pools

Indeed, in the set of wild accessions LD decayed within 2 cM on all chromosomes, except for chromosome 1, whether we corrected for GR or not. Since PCoA did not reveal any clusters of wild beets according to their geographical origin or described phenotypic type and due to the observation that LD decay did not change significantly after correction for GR, one can assume that the wild beets used within this study represent a completely unstructured set of accessions which reveal no apparent signatures of selection. Therefore, we concluded that the LD patterns in both of the elite heterotic pools reflect the breeding history of these elite lines, meaning that extensive artificial selection has been performed on these lines. Additionally, it is known that the genetic base of sugar beet is fairly narrow as this goes back to selections made by F.K. Achard in the late eighteenth century leading to the "White Silesian beet" which formed the basis of most modern varieties (Cooke and Scott 1993). However, it was unexpected for us to find no sign of population structure in the wild beet accessions since they originate from distant places all over Europe where they should have developed local adaptations that should be visible after a genetic diversity analysis such as the one we have performed. It is possible, but unlikely, that the geographical origins have been wrongly annotated leading to the present results. However, it could equally be that the number of different wild beet accessions used within this study is not big enough, and that we lack power to detect differences due to origin and phenotypic type. Moreover, we used a limited number of SNPs that had been designed to detect polymorphism between elite breeding lines, meaning that they might not be the best choice for a genetic diversity study on wild beet accessions. Nevertheless, gene diversity and heterozygosity statistics suggested a broader genetic basis in the wild beet accessions than in the elite breeding pools. This was expected as sugar beet is an outbreeding species while breeding lines, on the contrary, are strongly inbred with a narrow genetic basis. Large scale approaches to investigate the extent of genetic diversity, such as the French AKER program (www.aker-betterave.fr) with more than 2,000

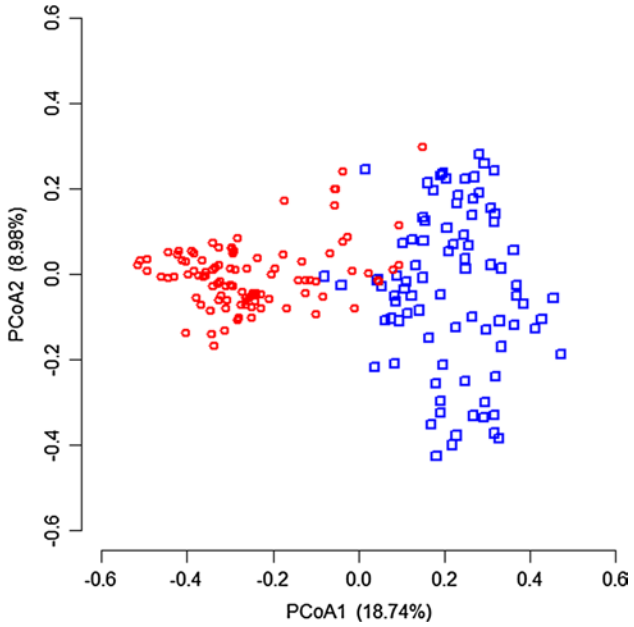

Fig. 6 Plot of the first two axes of the principal coordinates analysis using markers identified as significant by *Z*-test, excluding the SNPs on chromosome 3 and the SNPs associated with monogermy and maintainers of CMS on chromosome 4. The *red circles* indicate lines classified as pollen parent and the *blue squares* indicate lines classified as seed parent

individual accessions, are needed to study the genetic imprint of local adaptations in wild beets.

To our knowledge, this is the first study in sugar beet to provide a comparison of LD between SNPs with and without inclusion of the principal components capturing the genetic relatedness between elite breeding lines and wild beet accessions, a method previously described by Mangin et al. 2012. Our results indicate that the two elite heterotic pools were not equally structured, probably due to the more intensive selection in the sugar beet lines used as pollen parents when compared to the lines used as seed parents. Local LD decay plots (Fig. 3) revealed a region at approximately 82 cM on chromosome 4 which displayed strong LD with adjacent markers in the female elite lines, but not in the male heterotic pool. In the genome-wide pairwise F_{ST} scan, this region also showed high F_{ST}, suggesting divergence of the two pools (Fig. 5). Genetic mapping studies have located fruit monogermy in sugar beet exactly at that region on chromosome 4 (Barzen et al. 1992). Monogermy, which is inherited by a single recessive gene, is an important characteristic that was introduced into commercial varieties only in the second half of the twentieth century and is now part of all seed parent lines in commercial breeding programs (Cooke and Scott 1993). Additionally, this region on chromosome 4 also harbors the Z-locus which is, beside the X-locus on chromosome 3, responsible for the maintenance of cytoplasmic male sterility (Pillen et al. 1993; Schondelmaier and Jung 1997; Hagihara et al. 2005). As commercial sugar beet seed production relies on this CMS system (Owen 1945), maintainer alleles at both the X- and the Z-loci are present in all female breeding lines. Similarly, the resistance to rhizomania identified by the Holly Sugar Company in California (Lewellen et al. 1987) and conferred by the dominant *Rz1* locus has been widely introduced into almost all commercial sugar beet varieties since the early 1990s. *Rz1* has been mapped to a resistance gene cluster on chromosome 3 (Barzen et al. 1992; Scholten et al. 1999; Lein et al. 2007) and colocates with a region at 50 cM on chromosome 3 where we could detect adjacent markers which are in strong LD (Fig. 3) as well as reduced heterozygosity as indicated by the genome-wide calculation of F_{ST} (Fig. 5). The elevated levels of LD on chromosomes 3 and 4 (Fig. 3) point to the recent introduction of monogermy, the CMS maintainer loci and the *Rz1* locus into elite sugar beet germplasm. As expected, elevated LD around the monogermy locus was only observed in the seed parent lines, while LD decayed relatively quickly in the same region in the pollen parent pool. Strong LD was also detected around the *Rz1* locus on chromosome 3 in the male breeding lines, where it was deliberately introgressed, but not in the lines of the female heterotic pool, which for the most part do not contain this trait. The extent of LD around the monogermy and *Rz1* loci is also indicative of the time of introgression of these traits into the commercial sugar beet breeding lines. While strong local LD extends to an interval of approximately 10 cM around the monogermy locus, the elevated local LD persists in an interval of more than 20 cM at the region where the *Rz1* locus is located on chromosome 3. This is expected as monogermy was introduced in commercial sugar beet breeding lines starting from the 1950s onwards, while introgression of the *Rz1* locus started only in the 1990s (Cooke and Scott 1993), thereby allowing fewer cycles of recombination and less loss of potential linkage drag around the introgressed loci. Strong LD was also observed between markers located at the top of chromosome 9 in the female breeding pool (Fig. 3), and this region also exhibited high F_{ST} values (Fig. 5). While we can possibly explain the LD profiles observed on chromosomes 3 and 4 through selections that have been made by the breeders, we cannot provide such an explanation for the extensive LD observed on chromosome 9. Testing for significant allele frequency differences between the pollen and seed parent pools indicated that SNPs showing significant allele frequency differences were present on almost all chromosomes. Regions other than those that carry disease resistance loci or further genes involved in traits that have been actively under selection in one of both pools, such as monogermy, may contribute to the combining ability of the

two heterotic groups. Identification of such loci would be of particular importance within a breeding program, since this would reduce the number of testcrosses to be made and decrease the costs for assessing yield performance.

Differences between pollen and seed parent pools were also visible from the heat maps showing inter-chromosomal LD (Fig. 4a, b). LD between SNPs on different chromosomes was extensive in the pollen parent pool, and almost absent in the seed parent pool. Interestingly, inter-chromosomal LD observed in the pollen parent pool always involved loci on chromosome 3. The extensive inter-chromosomal LD reported for the pollen donor lines is possibly the result of selection for a specific combination of characteristics, among which is rhizomania resistance. The *Rz1* resistance to BNYVV comes from a single source and the introgression into elite sugar beet breeding lines occurred very rapidly. The strong selection pressure is highlighted as discussed above by the local LD decay plots (Fig. 3), but might have also created a genetic bottleneck leading to high LD on other regions of the genome, such as the strong LD on top of chromosome 9 and between loci on different chromosomes (Fig. 4).

Besides revealing selection history within and among genetic pools, the study of LD decay can be used to design future GWAS studies. Based on the LD decay estimates in the pollen and seed parent pools and the length of the chromosomes, it is possible to estimate the minimum number of SNPs required for successful GWAS. Since LD may decay at different rates in different genetic pool and chromosomes, as was the case in this study, it is advantageous to calculate the required number of markers per chromosome and per germplasm group. For example, from Table 2, for chromosome 1 in the pollen parent pool, GWAS would require at least one marker every 10.6 cM (=2 × 5.3), because this would guarantee markers to be always within 5.3 cM from a putative QTL. Note that the LD estimates in Table 2 are based on a 95 % quantile non-linear regression curve, which implies an upper bound for LD, which for the understanding of population genetic processes seems most informative. For assessing the minimum number of required markers for GWAS, using a lower bound for LD, e.g. by taking a 20 % quantile to estimate LD would be better. However, calculating such a lower quantile can be problematic if estimated positions of the SNPs on the linkage map are not precise enough, because there were not enough observed recombinations between closely linked markers in the mapping population. For example, for chromosome 1 in the pollen parent pool, the LD estimate is equal to zero if we use <40 % quantile non-linear regression (Suppl. Fig. 4) as the result of the underestimation of the distance between markers because of not enough map resolution. The uncertainty in the positions of the markers on the linkage map can be improved using bigger mapping populations.

We conclude that calculating LD per chromosome, with correction for genetic relatedness, is an interesting approach. On the one hand, it gives us a good estimate of the upper bound of LD, to get a better understanding of the population genetic processes. On the other hand, it also can be used as guideline in designing GWAS studies to define the number of markers needed.

References

Arnaud JF, Viard F, Delescluse M, Cuguen J (2003) Evidence for gene flow via seed dispersal from crop to wild relatives in *Beta vulgaris* (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. Proc R Soc Lond B 270:1565–1571

Arnaud J-F, Fénart S, Godé C et al (2009) Fine-scale geographical structure of genetic diversity in inland wild beet populations. Mol Ecol 18:3201–3215. doi:10.1111/j.1365-294X.2009.04279.x

Atwell S, Huang YS, Vilhjálmsson BJ et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631. doi:10.1038/nature08800

Barzen E, Mechelke W (1995) An extended map of the sugar beet genome containing RFLP and RAPD loci. Theor Appl Genet 90:189–193

Barzen E, Mechelke W, Ritter E (1992) RFLP markers for sugar beet breeding: chromosomal linkage maps and location of major genes for rhizomania resistance, monogermy and hypocotyl colour. Plant J 2:601–611

Clark RM, Linton E, Messing J, Doebley JF (2004) Pattern of diversity in the genomic region near the maize domestication gene tb1. Proc Natl Acad Sci USA 101:700–707

Cooke DA, Scott JE (1993) The sugar beet crop. Chapman and Hall, London

Desplanque B, Viard F, Bernard J et al (2000) The linkage disequilibrium between chloroplast DNA and mitochondrial DNA haplotypes in *Beta vulgaris* ssp. *maritima* (L.): the usefulness of both genomes for population genetic studies. Mol Ecol 9:141–154

Draycott AP (2006) Sugar beet. Blackwell Publishing Ltd, Oxford

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Res 10:564–567

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Pearson Education Limited, Essex

Fievet V, Touzet P, Arnaud JF, Cuguen J (2007) Spatial analysis of nuclear and cytoplasmic DNA diversity in wild sea beet (*Beta vulgaris* ssp. *maritima*) populations: do marine currents shape the genetic structure? Mol Ecol 16:1847–1864

Grimmer MK, Trybush S, Hanley S et al (2007) An anchored linkage map for sugar beet based on AFLP, SNP and RAPD markers and QTL mapping of a new source of resistance to *Beet necrotic yellow vein virus*. Theor Appl Genet 114:1151–1160. doi:10.1007/s00122-007-0507-3

Hagihara E, Itchoda N, Habu Y (2005) Molecular mapping of a fertility restorer gene for Owen cytoplasmic male sterility in sugar beet. Theor Appl Genet 111:250–255. doi:10.1007/s00122-005-2010-z

Jung C, Pillen K, Frese L (1993) Phylogenetic relationships between cultivated and wild species of the genusBeta revealed by DNA "fingerprinting". Theor Appl Genet 86:449–457

Kraft T, Fridlund B, Hjerdin A et al (1997) Estimating genetic variation in sugar beets and wild beets using pools of individuals. Genome 40:527–533

Kraft T, Hansen M, Nilsson NO (2000) Linkage disequilibrium and fingerprinting in sugar beet. Theor Appl Genet 101:323–326

Lein JC, Asbach K, Tian Y et al (2007) Resistance gene analogues are clustered on chromosome 3 of sugar beet and cosegregate with QTL for rhizomania resistance. Genome 50:61–71. doi:10.1139/G06-131

Lewellen RT, Skoyen JO, Erichsen AW (1987) Breeding sugar beet for resistance to rhizomania: evaluation of host-plant reactions and selection for and inheritance of resistance. Winter Congress of the IIRB. pp 139-156

Li J, Schulz B, Stich B (2010) Population structure and genetic diversity in elite sugar beet germplasm investigated with SSR markers. Euphytica 175:35–42. doi:10.1007/s10681-010-0161-8

Li J, Luhmann A-K, Weiszleder K, Stich B (2011) Genome-wide distribution of genetic diversity and linkage disequilibrium in elite sugar beet germplasm. BMC Genom 12:484. doi:10.1186/1471-2164-12-484

Liu K, Muse SV (2005) PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics 21(9):2128–2129

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 108:285–291. doi:10.1038/hdy.2011.73

McGrath JM, Derrico CA, Yu Y (1999) Genetic diversity in selected, historical US sugarbeet germplasm and *Beta vulgaris* ssp. *maritima*. Theor Appl Genet 98:968–976. doi:10.1007/s001220051157

Murray RS, Larry JS (2007) Theory and problem of statistics, 4th edn. Schaum's Outlines McGRAW-HILL, New York

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

Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD (2006) Selection under domestication: evidence for a sweep in the rice *waxy* genomic region. Genetics 173:975–983

Owen FV (1945) Cytoplasmically inherited male-sterility in sugar beets. J Agric Res 71:423–440

Palalsa K, Morgante M, Tingey S, Rafalski A (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. Proc Natl Acad Sci USA 101:9885–9890

Panella L, Lewellen RT (2007) Broadening the genetic base of sugar beet: introgression from wild relatives. Euphytica 154:383–400. doi:10.1007/s10681-006-9209-1

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLOS Genet 2:e190. doi:10.1371/journal.pgen.0020190

Pillen K, Sleinrucken G, Herrmann RG, Jung C (1993) An extended linkage map of sugar beet (*Beta vulgaris* L.) including nine putative lethal genes and the restorer gene X. Plant Breeding 111:265–272

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

Scholten OE, De Bock TS, Klein-Lankhorst RM, Lange W (1999) Inheritance of resistance to beet necrotic yellow vein virus in *Beta vulgaris* conferred by a second gene for resistance. Theor Appl Genet 99:740–746. doi:10.1007/s001220051292

Schondelmaier J, Jung C (1997) Chromosomal assignment of the nine linkage groups of sugar beet (*Beta vulgaris* L.) using primary trisomics. Theor Appl Genet 95:590–596. doi:10.1007/s001220050600

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

Smulders MJ, Esselink GD, Everaert I et al (2010) Characterisation of sugar beet (*Beta vulgaris* L. ssp. *vulgaris*) varieties using microsatellite markers. BMC Genet 11:41–52

Stich B, Melchinger AE, Heckenberger M et al (2008a) Association mapping in multiple segregating populations of sugar beet (*Beta vulgaris* L.). Theor Appl Genet 117:1167–1179. doi:10.1007/s00122-008-0854-8

Stich B, Piepho H-P, Schulz B, Melchinger AE (2008b) Multi-trait association mapping in sugar beet (*Beta vulgaris* L.). Theor Appl Genet 117:947–954. doi:10.1007/s00122-008-0834-z

VSN International (2011) Genstat for Windows 14th Edition

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

Würschum T, Maurer HP, Kraft T et al (2011a) Genome-wide association mapping of agronomic traits in sugar beet. Theor Appl Genet 123:1121–1131. doi:10.1007/s00122-011-1653-1

Würschum T, Maurer HP, Schulz B et al (2011b) Genome-wide association mapping reveals epistasis and genetic interaction networks in sugar beet. Theor Appl Genet 123:109–118. doi:10.1007/s00122-011-1570-3

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(12):e8451. doi:10.1371/journal.pone.0008451

Zhao K, Aranzana MJ, Kim S et al (2007) An *Arabidopsis* example of association mapping in structured samples. PLoS Genet 3:e4. doi:10.1371/journal.pgen.0030004