Theor Appl Genet (2015) 128:2255–2271
DOI 10.1007/s00122-015-2582-1

ORIGINAL ARTICLE

Breeding patterns and cultivated beets origins by genetic diversity and linkage disequilibrium analyses

Brigitte Mangin[1,2,3] · Florian Sandron[3] · Karine Henry[4] · Brigitte Devaux[4] · Glenda Willems[5] · Pierre Devaux[4] · Ellen Goudemand[4]

Received: 19 November 2014 / Accepted: 10 July 2015 / Published online: 4 August 2015

Abstract

Key message **Genetic diversity in worldwide population of beets is strongly affected by the domestication history, and the comparison of linkage disequilibrium in worldwide and elite populations highlights strong selection pressure.**

Abstract Genetic relationships and linkage disequilibrium (LD) were evaluated in a set of 2035 worldwide beet accessions and in another of 1338 elite sugar beet lines, using 320 and 769 single nucleotide polymorphisms, respectively. The structures of the populations were analyzed using four different approaches. Within the worldwide population, three of the methods gave a very coherent picture of the population structure. Fodder beet and sugar beet accessions were grouped together, separated from garden beets and sea beets, reflecting well the origins of beet domestication. The structure of the elite panel, however, was less stable between clustering methods, which was probably because of the high level of genetic mixing in breeding programs. For the linkage disequilibrium analysis, the usual measure (r^2) was used, and compared with others that correct for population structure and relatedness (r^2_S, r^2_V, r^2_{VS}). The LD as measured by r^2 persisted beyond 10 cM within the elite panel and fell below 0.1 after less than 2 cM in the worldwide population, for almost all chromosomes. With correction for relatedness, LD decreased under 0.1 by 1 cM for almost all chromosomes in both populations, except for chromosomes 3 and 9 within the elite panel. In these regions, the larger extent of LD could be explained by strong selection pressure.

Introduction

All cultivated beets belong to the same subspecies (*Beta vulgaris* L. *ssp vulgaris*), members of the genus *Beta* L. and of the family *Amaranthaceae* (formerly *Chenopodiaceae*), and are divided into four groups: leaf beets, garden beets, fodder beets and sugar beets (Gill and Vear 1958). The sugar beet crop supplies between a quarter and a third of the world's sugar, with sugar cane producing the remainder. While cane grows in tropical climates, sugar beet is best suited to continental climates, characterized by moderate temperatures and uniformly distributed rainfall. The major areas of sugar beet cultivation are in the US and in Europe (Russia, Ukraine, Germany, France…). The most important improvements in sugar beet breeding over the last 50 years have been the change in breeding methodologies with the introduction of hybrid varieties and marker-assisted selection (Biancardi et al. 2005), and the introduction of new traits in breeding schemes such as pest and disease resistances (rhizomania, sugar beet cyst nematode). The increase

Communicated by M. Frisch.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-015-2582-1) contains supplementary material, which is available to authorized users.

✉ Ellen Goudemand
ellen.goudemand@florimond-desprez.fr

1 INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, 31326 Castanet-Tolosan, France

2 CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, 31326 Castanet-Tolosan, France

3 INRA, Mathématique et Informatique Appliquées de Toulouse (MIAT), UR875, 31326 Castanet-Tolosan, France

4 S.A.S. Florimond-Desprez Veuve and Fils, BP41, 59242 Cappelle-en-Pévèle, France

5 SESVanderHave, Industriepark Soldatenplein Zone 2/Nr 15, 3300 Tienen, Belgium

of sugar yield per hectare in advanced European countries is about 1.4 % annually (Bosemark 2006), with approximately one half of this improvement attributed to breeding. All of today's commercial sugar beet varieties are hybrids, made possible by the discovery of cytoplasmic male sterility (CMS) (Owen 1945) which gave the possibility to develop male-sterile parental lines. Commonly, a maintainer ("O-type") line is hybridized with the male-sterile (CMS) equivalent of another line to produce a male-sterile F1. This (male-sterile) F1 will be used as the seed parent in crosses with pollinator lines (Bosemark 1993). The breeding for CMS lines and pollinator lines is done in separate breeding programs with regular interaction to evaluate the performance of the hybrids made between the lines.

The genetic diversity in cultivated sugar beet germplasm was found to be low compared with other open-pollinated crops (Bosemark 1979, 1989). Cultivated beets are likely to originate from wild sea beet (*Beta vulgaris* L. *ssp. maritima*) through different selection pressures (Cooke and Scott 1993). Wild beets (i.e., species and subspecies of the genus *Beta* outside *B. vulgaris ssp. vulgaris*) have only been used as potential sources of specific traits for cultivated beets, particularly disease resistance characters (Lewellen and Whitney 1993). Artificial hybridizations between species of the genus *Beta* and *Corollinae*, *Nanae* and *Procumbentes* genera have been tested but have proved very difficult (McGrath et al. 2007). Genes for nematode resistance have been introduced into sugar beet from wild beets of the section Procumbentes but monosomic addition or translocation lines, that were obtained, suffered from poor transmission rates of the resistance genes (Jung et al. 1994).

Agriculturally important traits are mostly controlled by quantitative trait loci (QTL). Genetic mapping and the identification of markers linked to these functional loci facilitate breeding for crop improvement. Linkage analysis, a method for dissecting complex traits, exploits the coinheritance of QTL and adjacent markers within biparental populations. Although based on the same fundamental principles of genetic recombination as linkage analysis, association mapping examines this shared inheritance for a collection of individuals with historical recombination. The comparatively high resolution provided by association mapping is dependent upon the extent of linkage disequilibrium (LD) across the genome (Zhu et al. 2008). Linkage disequilibrium (LD) is the nonindependence of alleles between loci (Hedrick 1987). LD will tend to decay with genetic distance along the chromosome, because genetically distant loci are more likely to have recombined in the past than tightly linked loci. LD can be the result of physical linkage of loci but it can also be due to nonrandom mating, recent mutation, genetic drift, natural or artificial selection. Factors affecting LD in plants have been extensively discussed in many papers (Flint-Garcia et al. 2003; Gaut and Long 2003; Gupta et al. 2005; Kim et al. 2007). Linkage disequilibrium between functional loci and markers that are physically linked is one of the key to association mapping success. On a technical level, it determines the marker density required for association mapping power. If LD decays rapidly, then a higher marker density is required to capture the effect of functional causal loci (Yu and Buckler 2006). Among the several LD measures, the most commonly used for diallelic loci is r^2, the Pearson squared correlation between the alleles at two different loci. Its user-friendliness and the proof by Pritchard and Przeworski (2001) that it is the measure linked to the resolution power of association mapping explains a great part of its success. However, population differentiation and relatedness present within the sample can lead to biased estimation of linkage disequilibrium and therefore to an inappropriate choice of marker density. Population structure due to inbreeding in many populations is a common source of error in LD mapping because the current measure of LD assumes random mating (i.e., without preferential crossings) between individuals belonging to the same community. Relatedness induces correlation between the observed marker genotypes, and in such situations the Pearson coefficient is no longer an unbiased estimator of the correlation of the loci alleles (McDonald 2014).

Three novel LD measures (Mangin et al. 2012) were derived from the usual r^2 measure taking into account the population structure (r^2_S), the population relatedness (r^2_V), or both (r^2_{VS}). These corrected r^2 measures actually captured 'true' LD, and were proved to be linked to the resolution power of association mapping when using the unified mixed model of Yu et al. (2005).

The extent of LD has been investigated in wild beet accessions (Adetunji et al. 2014). In this study, the LD of 99 wild beet accessions (20 *B. vulgaris ssp. maritima* and 79 *B. vulgaris ssp. vulgaris*), covering a wide range of beet types and geographical origin, was investigated with 498 SNPs. Adetunji et al. (2014) reported that LD decayed within a map distance of about 1.7 cM without genetic relatedness correction and within 1.05 cM with correction. In addition, LD was also analyzed many times in elite sugar beet lines (Kraft et al. 2000; Li et al. 2010, 2011; Adetunji et al. 2014). Li et al. (2011) examined a total of 502 sugar beet inbreds from the pollen parent heterotic pool with 328 SNPs and reported significant LD between pairs of loci that were 7 cM apart. Adetunji et al. (2014) observed LD in 234 elite sugar beet breeding lines, 139 belonging to the pollen parent pool and 95 to the seed parent pool, with the same 498 SNP markers used for their analysis in the wild beet population. When genetic relatedness was not accounted for, LD extended beyond 50 cM on four chromosomes in the pollen parent pool and three chromosomes in the seed parent pool. When genetic relatedness was corrected for,

LD decayed within a distance of less than 6 and 4 cM on all chromosomes in the pollen and seed parent pools, respectively. Comparing the extent of LD, with markers that can be located on the same genetic map, in two different panels composed of breeding lines and wild beets, respectively, allowed to highlight breeding patterns caused by high selection pressure in some regions of the genome.

Structure of a population is an important feature that can be known when the population is sampled within different geographical regions but is usually unknown when the population is just a collection of individuals. However, population structures have a great impact on the results of association studies as shown by Mezmouk et al. (2011) and can mislead LD interpretation when the usual r^2 measure is used by showing long-range extent (Mangin et al. 2012) as well as dependencies between unlinked loci. The most common methods used to infer population structures are Patterson's (2006) PCA and the software application STRUCTURE (Pritchard et al. 2000). They have been recently put in a unified framework (Engelhardt and Stephens 2010) and compared with simulated and real human populations. When the set of individuals are descendants of few well-differentiated ancestral populations, the two methods give essentially the same results even if the loadings of STRUCTURE are more interpretable. In more complex situations, the generality of the above conclusion must be investigated.

In the present study, population structure and linkage disequilibrium were analyzed in one set of 2035 worldwide beet accessions and in one set of 1338 elite sugar beet lines, with 320 and 769 SNPs, respectively, after filtering on minor allele frequency (MAF) and missing data. The worldwide beets set was made up of accessions belonging to various species and subspecies of the genus *Beta*, and the elite panel was composed of lines that are used as pollinators in breeding programs.

Materials and methods

Plant material

A total of 1367 elite lines issued from Florimond-Desprez's proprietarybreeding pool and a worldwide population of 2246 accessions were used in this study before data cleaning.

The 2246 accessions were collected by agricultural research institutes around the world and maintained in different gene banks. Information on the accessions (mainly country of origin, species and subspecies) were found in web databases such as GRIN-ARS (USDA, US, http://www.ars-grin.gov/npgs/) and EURISCO (ECPGR, Europe, http://www.ecpgr.cgiar.org/germplasm_databases.html). The majority of the accessions were maintained in 3 international gene banks in US, Germany and Greece. This worldwide set included wild accessions from gene pool I of genus Beta and cultivated beet accessions (see Table 1). It covered a broad range of geographical origins. We tried, where possible, to cover all of the native origins of wild beets and of growing regions for the cultivated accessions. The collecting sites reported for wild beets ranged from Atlantic coast of Morocco to Denmark, and along the Mediterranean coast from Algeria to Israel and Turkey. For cultivated beets, most regions were represented, including Western Europe, Southern Europe, Central and Eastern Europe, Asia, the Middle East, the US.

Table 1 Composition of the worldwide population according to species information

Species	Number of accessions
Beta[a]	15
Beta macrocarpa	30
Beta vulgaris[b]	143
Beta vulgaris ssp.adanensis	10
Beta vulgaris ssp.maritima	997
Betavulgaris ssp.vulgaris[c]	78
Beta vulgaris ssp.vulgaris cultigroup fodder beet	213
Beta vulgaris ssp.vulgaris cultigroup garden beet	293
Beta vulgaris ssp.vulgaris cultigroup sugar beet	420
Beta vulgaris ssp.vulgaris cultigroup leaf beet/swiss chards type[d]	47

[a] Accessions classified as *Beta* but no species specified

[b] Accessions classified as *Beta vulgaris* but the subspecies is unknown

[c] Accessions classified as *Beta vulgaris ssp.vulgaris* but the cultigroup is unspecified

[d] Accessions representing only the swiss chards type in the leaf beet cultigroup. This cultigroup is originally composed of 2 beets types, spinach beets and swiss chards

The elite sugar beet breeding pool is composed of 1367 proprietary lines, representing the pollinator pool that is used in breeding programs of Florimond Desprez.

SNP markers and genotyping

The SNP makers used in this study were designed in both genic and intergenic sequences (cDNAs) in a set of elite lines and had previously been mapped using three different F2 mapping populations, as described by Adetunji et al. (2014). The length of the total genetic map is 705 cM, with individual chromosome sizes estimating between 70 cM for chromosome 5 and 91 cM for chromosome 3.

The samples used for DNA fingerprinting profiles were leaves of one plant per accession or breeding line. Leaf disks were sampled, frozen at −80 °C and freeze-dried. DNA extraction was performed using the NucleoSpin® Plant kit (Machery-Nagel, Düren, Germany) and genotyping was performed for individual SNPs using KASP genotyping chemistry (LGC Genomics, Teddington Middlesex, UK).

For genotyping elite material, a total of 836 SNPs were used. A more restricted set of markers containing 377 SNPs were used for genotyping of gene bank accessions. The selection was based on the level of polymorphism and to obtain a uniform coverage of the sugar beet genome.

Population structure

Several methods were applied to study the subpopulation structure in the worldwide population and the elite panel. We used principal component analysis (PCA) proposed by Patterson et al. (2006), which is a classical centered and scale PCA where the scale factors are computed assuming the Hardy–Weinberg equilibrium. To deal with missing observations, we used the package missMDA implemented in R (http://cran.r-project.org/web/packages/missMDA/index.html). The function imputes PCA of the above package handles missing data by a multiple imputation method that does not have the bias of single imputation methods. The well-used mean imputation method is known, for example, to reduce the variability of the data and to distort their correlation (Josse et al. 2011). As proposed by Patterson et al. (2006), the Tracy–Widom test (Tracy and Widom 1994) was applied to the PCA eigenvalues and to decipher the number of subpopulations as the number of significant eigenvalues plus 1. We also used the software application STRUCTURE (Pritchard et al. 2000) and the Evanno (2005) criterion to decipher the number of subpopulations. The results were obtained on Structure Harvester web site (Earl and vonHoldt 2012) (http://taylor0.biology.ucla.edu/structureHarvester/). We ran STRUCTURE for an admixture model with independent allele frequency between subpopulations. We varied the number of subpopulations K from 1 to 10 and we made five replicates for each K. The simulated MCMC chain had 150,000 iterations with the first 50,000 used as the burning period. We also investigated community algorithms implemented in the R package igraph (Csardi and Nepusz 2006) on the graph obtained using the nearly 10 % higher values of the alike-in-state (AIS) kinship matrix. Nodes of the graph were the accessions of the worldwide population or the genotypes of the elite panel. An edge was present between two nodes if and only if the AIS kinship coefficient between the corresponding individuals was superior to the mean plus 1.64 times the standard deviation of the AIS kinship coefficients in the population. The AIS kinship matrix was computed using the allele-matching kernel and the pairwise complete observations. Denoting $G_1$and G_2, the (0, 1, 2) coded genotype at each biallelic marker of the individual 1 and 2, respectively, where 0 stands for homozygote for the first allele, 1 for heterozygote and 2 for homozygote for the second allele, and assuming that the genotypic data do not have missing observations, the allele-matching kernel was computed by:

$$\frac{G_1^t G_2 + (2 - G_1)^t (2 - G_2)}{4M_{12}},$$

where M_{12} is the number of completely informative markers between individual 1 and 2. With igraph, we ran the fast greedy (Clauset et al. 2004), the leading eigenvector (Newman 2006), the multilevel (Blondel et al. 2008), the label propagation (Raghavan et al. 2007), the walktrap (Pons and Latapy 2006) and the infomap (Rosvall et al. 2009) algorithms. The first three methods try to optimize the modularity. Fast greedy is working bottom-up by merging communities although leading eigenvector is working top-down by splitting communities. Multi-level uses two fast greedy steps, the first at the vertex level then at the community level. The label propagation propagates the most frequent label among neighborhoods. Both walktrap and infomap are based on a random walk on the graph. The former merges communities depending on the results of short walks, the latter tries to optimize the shortest description length. Among them, we chose the assignation of the individuals by the algorithm reaching the maximum of modularity. We finally run a hierarchical clustering method with a dissimilarity based on the AIS kinship matrix using the R function hclust and its ward method.

A k-means algorithm on the significant axes of the PCA was used to assign each individual to a subpopulation. The R function kmean was used for this purpose. A simple threshold method that assigns an accession to a sub-population if and only if the maximum of the admixture proportions given by STRUCTURE is superior to a threshold was used on the STRUCTURE results. The community algorithms have an initial goal of assignation to a community so

Table 2 cM length covered, number of SNPs and number of different loci (same position on the genetic map) of each chromosome for the two populations compared to the length of the reference map

Chromosome	Reference map	Worldwide population			Elite panel		
	Length	Length	# SNP	# Loci	Length	# SNP	# Loci
1	74	66	18	18	66	88	33
2	74	74	32	30	61	63	21
3	91	91	69	48	91	172	50
4	86	83	50	40	83	122	39
5	70	68	37	30	68	116	39
6	79	79	37	23	70	47	21
7	79	66	28	22	64	41	20
8	78	72	32	27	77	71	31
9	74	65	16	15	69	49	24
Mean	78.3	73.8	35.4	28.1	72.1	84.4	30.9
Total	705	664	319	253	649	769	278

do not need a supplementary step. We cut the hierarchical tree to obtain the number of subpopulations found by the Tracy–Widom test using the R function cutree.

Principal component analysis plots were obtained by the plot3dIndiv function of the mixOmics R package (http://cran.r-project.org/web/packages/mixOmics/index.html), and Venn diagram showing the overlap between clustering methods was obtained with the VennDiagram R package (http://cran.r-project.org/web/packages/VennDiagram/index.html).

Linkage disequilibrium analysis

Linkage disequilibrium (LD) was estimated by the measures that correct for the population structure and relatedness implemented in the R package LDcorSV (Mangin et al. 2012; http://cran.r-project.org/web/packages/LDcorSV/index.html). The genotypic data were first imputed using fastphase v1.4 (Scheet and Stephens 2006). This allowed computing the LD measures using a new option of the LDcorSV package that avoids inverting the kinship matrix for each marker couple and so performs a quicker computation. We used two structure matrices when computed the LD measure corrected for the population structure. One was a 0–1 matrix obtained by the k-means algorithm on the PCA for the number of subpopulations found by the Tracy–Widom test. The second was the admixture proportions provided by STRUCTURE with a number of subpopulations found by the Evanno criterion.

We also used two variance–covariance matrices to correct the LD measure for relatedness. The first one was the AIS kinship matrix; it was used to get an unbiased estimator of the squared correlation between two markers. The second was equal to $I + K_{\mathrm{AIS}}$ where I denotes the identity matrix and K_{AIS} the AIS kinship matrix. It was used to be linked to the power of the association mixed-model test of Yu et al. (2005) for a trait with a heritability of 0.5 (Mangin et al. 2012).

The decay of LD was estimated with the model for linkage decay proposed by Hill and Weir (1988):

$$\mathrm{Esp}\left(r^2\right) = \left(\frac{10+c}{(2+c)(11+c)}\right)\left(1+\frac{(3+c)\left(12+12c+c^2\right)}{N(2+c)(11+c)}\right)$$

with $c = 4N_e d$. Where N is the sample size, N_e is the effective population size and d the distance in centimorgan between the two loci.

We also computed the proportion of genome that was covered by consecutive markers in LD when the LD measure was greater than a threshold. This proportion is an estimation of the genomic coverage of an association test for a given minimum power. When several markers were in the same locus, the maximum of LD measures with other consecutive markers were taken.

We did all usual statistics and our own computational developments on R (R Core Team 2014).

Results

Genotypic data

The genotypic data were filtered to keep markers with a MAF superior to 0.005 with less than 10 % of missing data and individuals with less than 10 % of missing data. This led to keep 2035 accessions for the worldwide population and 1338 genotypes for the elite panel with 320 and 769 SNPs, respectively. The coverage of the SNPs on the nine chromosomes of the sugar beet genome are presented in Table 2 (one marker of the worldwide population was not located on the map). The standard nomenclature of the nine chromosomes of sugar beet (Butterfass 1964) is used.

The marker MAF distribution was quite different between the worldwide population and the elite panel with a mean equal to 0.30 and 0.22, respectively, which reflects a lack of SNPs with rare alleles in the worldwide population compared to the elite panel that presents an inflation of SNPs with rare alleles (see Fig. S1 and Fig. S2 in the Supplementary data).

Population structure

As expected, the relatedness measured by the AIS kinship coefficients was lower in the worldwide population than within the elite panel with a mean equal to 0.61 and 0.70, respectively. However, the dispersion of the above coefficients was similar in the two populations with a standard deviation of 0.049 and 0.052, respectively. The distribution of the AIS kinship coefficients was skewed for the worldwide population and symmetric for the elite panel (see Fig. S3 and Fig. S4 in the Supplementary data).

The Tracy–Widom test performed on the PCA matrix proposed by Patterson et al. (2006) gave four and six significant eigenvalues for the worldwide population and the elite panel, respectively. We thus concluded to five and seven subpopulations, respectively. The six significant axes of the elite panel explained 8.65, 6.89, 6.84, 4.31, 4.11 and 3.83 % of the total variability. The percentages were more contrasted with 10.69, 5.76, 4.10 and 2.76 % for the four significant axes in the worldwide population. However, the elite panel was genotyped with nearly 60 % more SNPs compared to the worldwide population which could explain a smoother eigenvalue distribution in the elite panel.

The Evanno criterion gave four and six subpopulations for the worldwide population and the elite panel, respectively (see Fig. S5 and Fig. S6 in the Supplementary data). In both populations, a large part of accessions was assigned as admixed by STRUCTURE so the 0.65 threshold method managed to classify only 67 and 44 % of the worldwide population and the elite panel, respectively. As expected the accessions within the worldwide population were less admixed than within the elite panel.

A large part of the worldwide population was not included in the community study since 601 among 2035 accessions did not have large enough AIS kinship coefficients. This was completely different for the elite panel with only 24 among 1338 genotypes isolated by the filter on AIS kinship coefficients. For both populations, the optimum community algorithm was the multi-level algorithm. It clustered the worldwide population and the elite panel in 40 and 13 communities of at least two members, respectively. Six and seven large communities (>50 members) were found for the worldwide population and the elite panel, respectively. The community algorithm did not create new isolated individuals in either population.

Figure 1 shows the classes formed by the four clustering methods on the first three axes of the Patterson's PCA for the worldwide population. Figure 2 presents the same plots for the elite panel. Only large communities of more than 50 members were colored for the community method. Apart from the community method, the results were clear within the worldwide population, which was made up of four extreme groups (the cyan, the red, the green and the yellow colored groups). The central group (blue one) was composed of admixed accessions for STRUCTURE and was put in the same group for both the k-means method on the PCA and the hierarchical clustering method on the AIS kinship matrix since they do not model admixed individuals. The community method had a very different behavior. It found two extreme groups (the cyan and the green one) but it extended the yellow colored group, it cut the blue central group into two parts and found mostly small communities in the extreme red group. Pictures were less clear for the elite panel. Apart from the two extreme groups (red and green) that were present for the four methods but that differed largely in size, all other groups were split or joined depending on the methods.

The worldwide population was mostly composed (91 %) of *B. vulgaris ssp. vulgaris* and *B. vulgaris ssp. maritima.* Table 3 presents how those subspecies of *B. vulgaris* were classified by the clustering methods, except for the community method that gave a slightly different picture of the population structure. The three clustering methods gave mostly a similar classification of the subspecies. The red and green groups were mostly composed of *B. vulgaris ssp. vulgaris* and the cyan and yellow group of *B. vulgaris ssp. maritima.* The last group was a nearly balanced mixture of the two subspecies. Figure 3 shows the overlap between the clusterings of the worldwide accessions realized with the three different methods. In total, 1540 accessions out of 2035 were classified in the same groups by the three clustering methods. The k-means algorithm on the significant axes of the PCA and the hierarchical clustering on the AIS matrix seemed to be the closest clustering methods with 1819 accessions clustered in the same groups. Contrary to k-means algorithm and the hierarchical clustering on the AIS matrix, the clustering of STRUCTURE leaned to classify more *B. vulgaris ssp. maritima* into the cyan group and less into the central blue one.

Among the worldwide population, 882 accessions of *B. vulgaris ssp. vulgaris* had a phenotypic-"type" description (e.g., leaf beet, garden beet, fodder beet and sugar beet). However, the leaf beet description was visually confirmed only for the swiss chard type, so we limited our classification to garden beet, sugar beet, fodder beet and swiss chard. Table 4 presents how these accessions were classified by the k-means algorithm on the PCA matrix and the AIS matrix with a hierarchical clustering in five

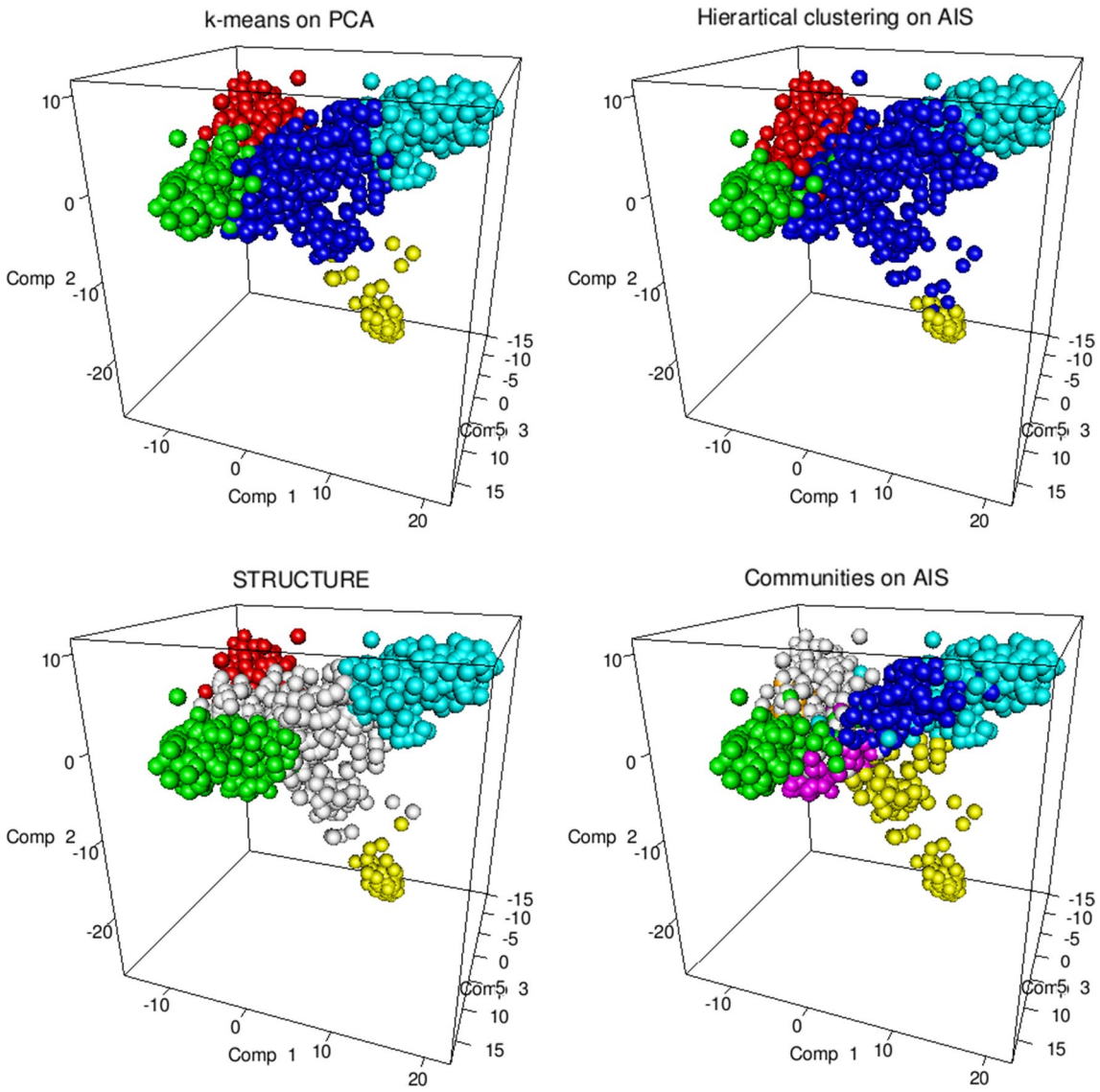

Fig. 1 Plots of the first three axes of Patterson's principal component analysis colored by four different clustering methods within the worldwide population. *Top-left* k-means clustering on the significant axes of Patterson's principal component analysis. *Top-right* cut of the hierarchical clustering tree obtained using the alike-in-state kinship matrix. *Bottom-left* clustering using the admixture proportions provided by STRUCTURE. *Bottom-right* community method on a graph with accessions linked when having high alike-in-state kinship coefficients

subpopulations. The two classifications were in good agreement. Each type of accession clustered mostly within a single class. Moreover, each type of accession had its own class except sugar and fodder beets that clustered in the same class. Sugar beet and swiss chard accessions were mainly classified into the red and the blue groups, respectively. Fodder beets were mostly clustered into the red group, with sugar beets, but they were also into the green and blue groups. Finally, garden beets were mostly present into the green group but some of them were classified into the red and blue groups. The k-means algorithm classified less fodder beet accessions into the red group (123) than the hierarchical clustering on the AIS matrix (166), but on the opposite it clustered more garden beet accessions into the green group (217/188). The Pearson Chi^2 test of independence between the annotation and the classification had a p value less than 10^{-15}. This illustrates that there was a clear genetic structure within the beet accessions since clusters could be described significantly by the phenotypic-type description.

The other classifications (threshold on STRUCTURE admixture proportions or communities) gave similar but less convincing results since some accessions were not classified (data not shown).

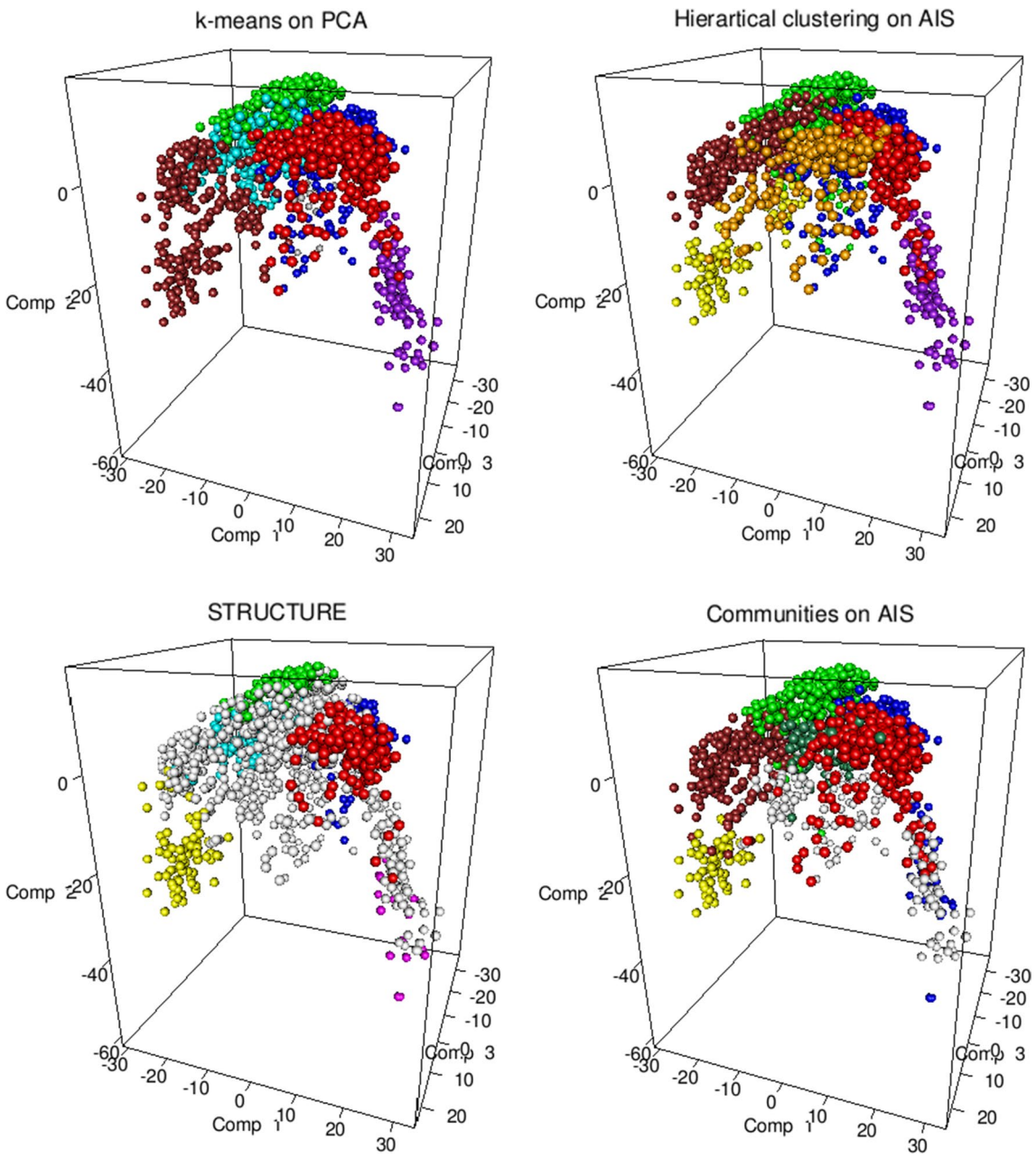

Fig. 2 Plots of the first three axes of Patterson's principal component analysis colored by four different clustering methods within the elite panel. *Top-left* k-means clustering on the significant axes of Patterson's principal component analysis. *Top-right* cut of the hierarchical clustering tree obtained using the alike-in-state kinship matrix. *Bottom-left* clustering using the admixture proportions provided by STRUCTURE. *Bottom-right* community method on a graph with accessions linked when having high alike-in-state kinship coefficients

Linkage disequilibrium (LD)

The LD is largely corrected by the relatedness kinship matrix alone, as illustrated for chromosome 3 in Fig. 4 within the elite panel. The six heat maps plot the LD as measured by the usual r^2 measure, the measure corrected for the relatedness computed with the AIS kinship matrix r^2_V, the measure corrected for the population structure r^2_S for two structure matrices, one computed with k-means on PCA and the other provided by STRUCTURE, and finally the measure corrected for both relatedness and population structure r^2_{VS}. The corrections for population structure gave nearly the same level of LD whichever population structure matrix we used. Despite that the correction for the population structure set to zero the long-ranged r^2 level as shown by the r^2_S measure plot, the effect of the population structure correction after the correction using the AIS kinship matrix seemed negligible. This was true for all the chromosomes

Table 3 Count table using the subspecies *vulgaris* or *maritima* of *B. vulgaris* accessions and the classes obtained by the k-means algorithm on the PCA matrix, the hierarchical clustering on the AIS matrix and the threshold on STRUCTURE admixture proportions

Class	1	2	3	4	5
Color	Red	Cyan	Green	Yellow	Blue/white
Ssp.	k-means on PCA				
Maritima	26	374	10	110	384
Vulgaris	514	5	283	0	151
Ssp.	Hierarchical clustering on AIS				
Maritima	22	324	6	99	453
Vulgaris	592	4	201	0	151
Ssp.	STRUCTURE				
Maritima	19	412	38	103	332
Vulgaris	417	6	297	0	233

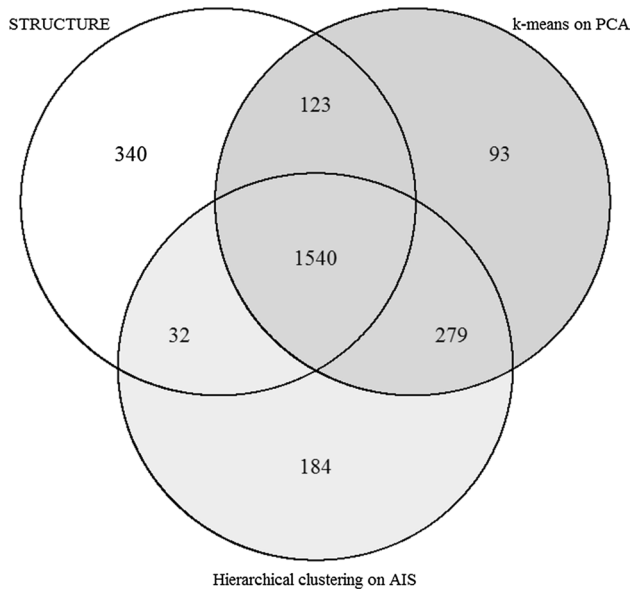

Fig. 3 Venn diagram showing the overlap between the clusterings of the worldwide accessions realized with k-means on the significant axes of Patterson's principal component analysis, the hierarchical clustering on the alike-in-state kinship matrix, and the admixture proportions provided by STRUCTURE

(see Fig. S7 to Fig. S14 in the Supplementary data). Figure 5 presents the same plots within the worldwide population. With a much lower marker density, the corrections of the usual r^2 measure were less impressive. However, the same conclusions were still valid, so the following results used the r^2_V measure of LD as the corrected one. This was also true for all the chromosomes (see Fig. S15 to Fig. S22 in the Supplementary data).

Some regions exhibit different LD level between the two populations that were visible despite their different marker densities. This was particularly the case of chromosome 3, for which the elite panel had a region of LD between 46 and 56 cM on the genetic map that was clearly absent in the worldwide population. This high level of LD on chromosome 3 coincides with a region that is known to be under strong selection pressure in elite sugar beet germplasm. This chromosome is known to harbor rhizomania resistance genes that have been strongly selected in elite germplasm (Biancardi et al. 2002; Scholten et al. 1999).

A global comparison of LD between the worldwide population and the elite panel is shown in Fig. 6. The Hill and Weir (1988) model of LD decay was adjusted and plotted for each chromosome with the r^2 and r^2_V measures. Clearly the LD level was largely overestimated when measured with r^2 within the elite panel. For almost all chromosomes, the correction brought by the AIS kinship matrix was sufficient to get a LD close to that observed within the worldwide population. Chromosomes 1 and 2 were remarkable with nearly superposed curves of LD decay when corrected. For the other chromosomes, the LD level was slightly higher in the elite panel than in the worldwide population. The highest differences were seen in chromosomes 3 and 9.

Mangin et al. (2012) proved that to achieve the same power at a marker linked to a causal locus with the test of association in the mixed model of Yu et al. (2005), the number of genotyped individual has to be increased by an inflation factor of $\frac{1}{r^2_{VS}}$, where V is the variance–covariance matrix of the trait in the above model. This V matrix is proportional to $I + \frac{h^2}{(1-h^2)}K$, where h^2 denotes the trait heritability. Let's suppose that a causal locus has a strong

Table 4 Count table using the cultigroup of *B. vulgaris ssp. vulgaris* accessions and the classes obtained by the k-means algorithm on the PCA matrix and the hierarchical clustering on the AIS matrix

	k-means on PCA					Hierarchical clustering on AIS				
Class	1	2	3	4	5	1	2	3	4	5
Color	Red	Cyan	Green	Yellow	Blue	Red	Cyan	Green	Yellow	Blue
Cultigroup										
Sugar beet	343	3	8	0	14	346	3	4	0	15
Fodder beet	123	0	52	0	27	166	0	7	0	29
Garden beet	13	0	217	0	40	35	0	188	0	45
Swiss chard	0	0	0	0	42	3	3	0	0	39

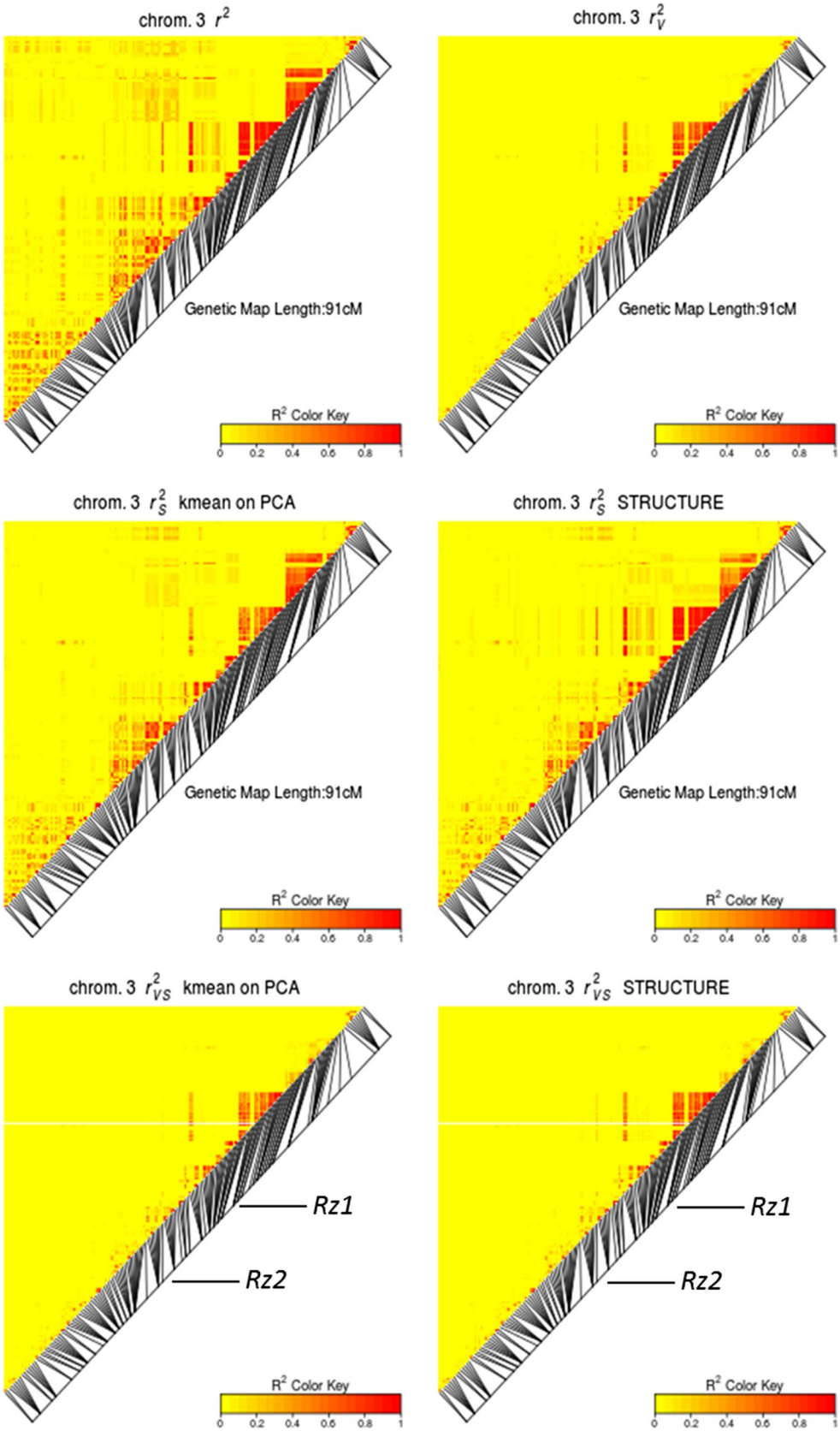

Fig. 4 Heat maps of linkage disequilibrium within the elite panel on chromosome 3. *Top-left* r^2 measure. *Top-right* measure corrected for the relatedness r_V^2 computed with the alike-in-state kinship matrix. *Middle-left* measure corrected for the population structure r_S^2 with structure matrix obtained with k-means on significant axes of Patterson's principal component analysis. *Middle-right* r_S^2 using the admixture proportions provided by STRUCTURE. *Bottom-left* measure corrected for both the relatedness and the population structure r_{VS}^2 with structure matrix obtained with k-means on significant axes of Patterson's principal component analysis. *Bottom-right* r_{VS}^2 using the admixture proportions provided by STRUCTURE. *Rz1* and *Rz2* correspond to the genetic mapping location of both BNYVV resistance genes

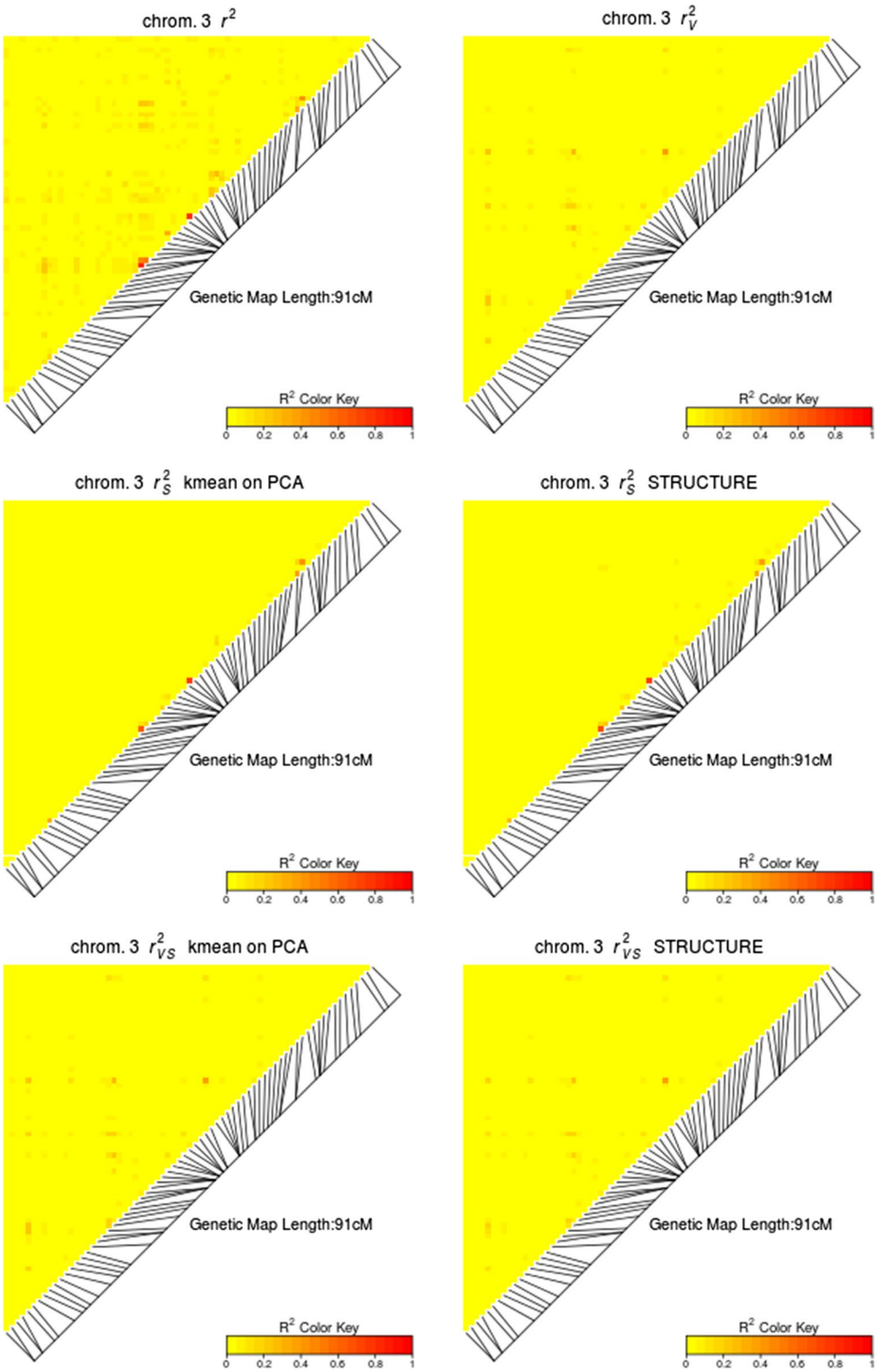

Fig. 5 Heat maps of linkage disequilibrium within the worldwide population on chromosome 3. *Top-left* r^2 measure. *Top-right* measure corrected for the relatedness r_V^2 computed with the alike-in-state kinship matrix. *Middle-left* measure corrected for the population structure r_S^2 with structure matrix obtained with k-means on significant axes of Patterson's principal component analysis. *Middle-right* r_S^2 using the admixture proportions provided by STRUCTURE. *Bottom-left* measure corrected for both the relatedness and the population structure r_{VS}^2 with structure matrix obtained with k-means on significant axes of Patterson's principal component analysis. *Bottom-right* r_{VS}^2 using the admixture proportions provided by STRUCTURE

effect and that it can be detected with a marker in LD equal to 0.1, if the LD between consecutive markers is assumed nearly constant in expectation within the interval they flanked, then this causal locus will be detected if it is located within intervals with flanking markers in LD greater or equal to 0.1. This is why we studied the proportion of genome covered by consecutive markers in LD greater or equal to a given LD threshold. These

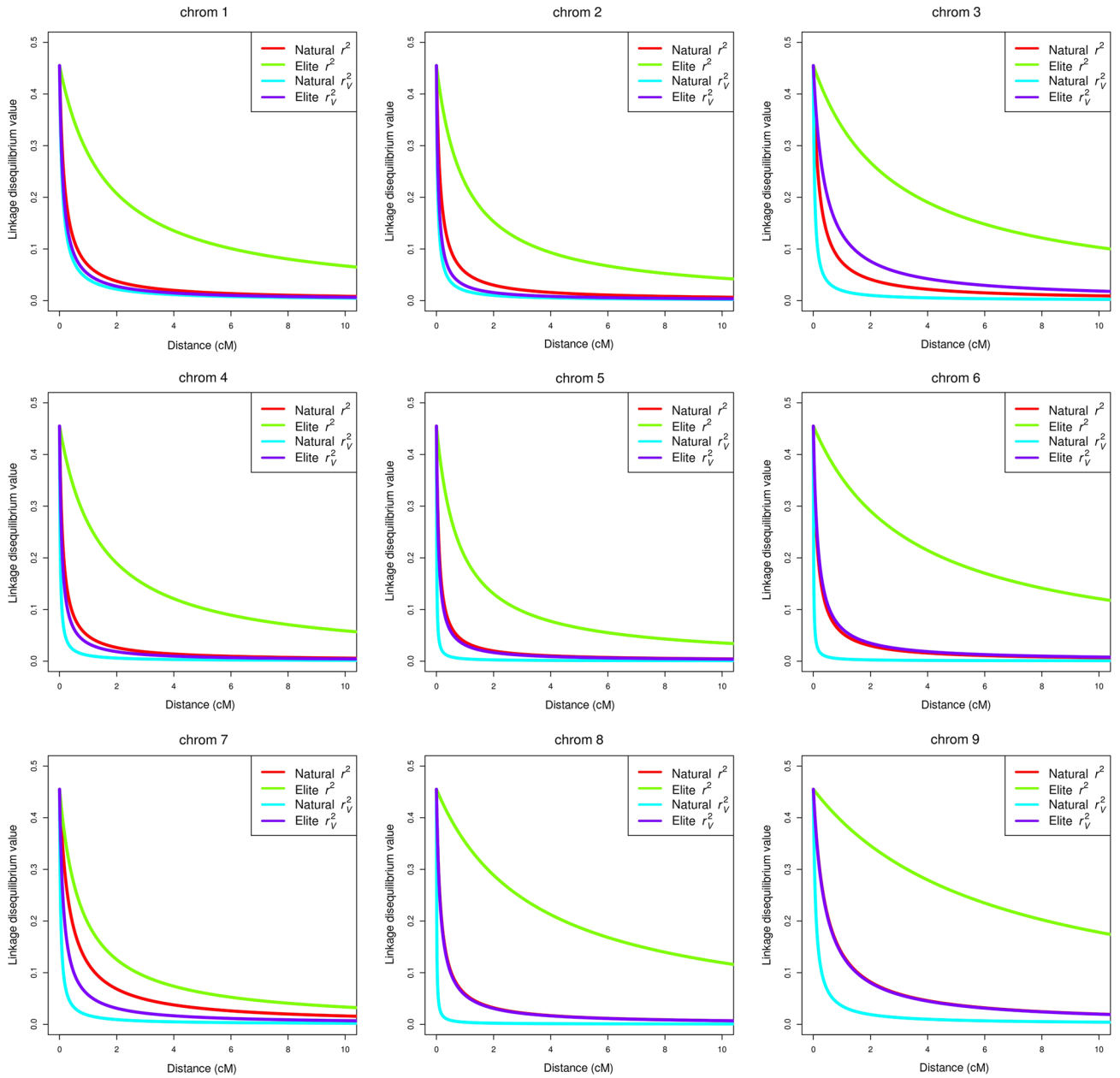

Fig. 6 Linkage disequilibrium decay adjusted by the Hill and Weir model (1988) for each of the nine chromosomes. r^2 within the worldwide population (*red*), r^2 within the elite panel (*green*), measure corrected for the relatedness computed with the alike-in-state kinship matrix r^2_V within the worldwide population (*blue*), r^2_V within the elite panel (*purple*)

proportions of genome coverageare presented in Table 5 for the elite panel and the worldwide population by chromosome and for the whole genome. We set the LD thresholds to 0.1, 0.2, and 0.3. We show the coverage for three different variance–covariance matrices: the identity matrix which is equivalent to compute the r^2_S measure and could correspond to very low heritability trait, $I + K_{\mathrm{AIS}}$ which corresponds to a trait of middle heritability and K_{AIS} which could correspond to a high heritability trait (when h^2 tends toward 1, the V matrix tends to be proportional to K). In all the cases, the STRUCTURE admixture proportions were used to correct for fixed structure effects to be in the so-called "$Q + K$" association model of Yu et al. (2005).

Table 5 Proportion of genomic coverage of consecutive markers in LD superior to a threshold for the elite panel and the worldwide population by chromosome and for the whole genomes

	Elite Panel									Worldwide population		
Var. Cov.	I			$I + K_{AIS}$			K_{AIS}			I	$I + K_{AIS}$	K_{AIS}
Threshold	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1[a]	0.1	0.1
Chromosome												
1	0.71	0.39	0.27	0.59	0.24	0.21	0.24	0.12	0.12	0.04	0.00	0.00
2	0.93	0.60	0.57	0.71	0.54	0.49	0.49	0.32	0.21	0.08	0.00	0.00
3	0.50	0.28	0.17	0.28	0.19	0.09	0.17	0.05	0.05	0.05	0.01	0.03
4	0.86	0.65	0.60	0.68	0.57	0.42	0.43	0.30	0.21	0.06	0.00	0.03
5	0.56	0.33	0.33	0.33	0.33	0.14	0.13	0.07	0.03	0.01	0.00	0.00
6	0.78	0.51	0.30	0.66	0.22	0.12	0.11	0.07	0.02	0.03	0.02	0.00
7	0.61	0.52	0.38	0.42	0.30	0.19	0.18	0.09	0.03	0.00	0.00	0.00
8	0.62	0.51	0.32	0.57	0.25	0.19	0.19	0.12	0.06	0.06	0.00	0.00
9	0.48	0.33	0.07	0.31	0.07	0.05	0.07	0.03	0.03	0.00	0.00	0.01
Whole	0.68	0.47	0.35	0.52	0.31	0.22	0.23	0.14	0.09	0.04	0.00	0.01

Three variance–covariance matrices were used to compute the r^2_{VS} measure and we set three thresholds. The S matrix in r^2_{VS} computation was the STRUCTURE admixture proportions

[a] Results for the other thresholds were all close to 0

Discussion

We have performed a study of population structure and linkage disequilibrium with two different populations of beet accessions. The worldwide population, formed of 2246 accessions from all over the world, was genotyped with 377 SNPs. The elite panel comprised 1367 lines, constituted part of the proprietary breeding program of Florimond-Desprez, and was genotyped with 836 SNPs.

Population structure

We conducted the population structure analysis with four different approaches: Patterson's (2006) PCA, the software application STRUCTURE (Pritchard et al. 2000), a hierarchical clustering on a kinship matrix and a community analysis using a graph built on kinship coefficients. We showed that within the worldwide population, the three first methods gave a very coherent picture of the population structure. The worldwide population was composed of four extreme populations and accessions that appear to be admixed, on the basis of the STRUCTURE proportion results. Two of the four extreme populations contained mostly *B. vulgaris ssp. vulgaris* with sugar and fodder beet in one population and garden beet in the other. *B. vulgaris ssp. vulgaris* for leaf was mostly made up of admixed accessions. A third extreme group comprised mostly *B. vulgaris ssp. maritima.* We did not identify a common characteristic for members of the last extreme group. This structure could be explained by the origin of beet domestication. The oldest known cultivated beet type was chard (leaf beet), domesticated from sea beets (*Beta vulgaris ssp. maritima*) and grown by Greeks and Romans for medicinally purposes and for their dense foliage. All other cultivated groups (garden, fodder and sugar) are originated from leaf beets with expanded hypocotyl and root. This ancient origin of leaf beets could partially explain that swiss chards accessions were only found in the mixture class with accessions coming from other cultigroups. Moreover, the first modern sugar beet came from selections made in the middle of the eighteenth century from white fodder beets grown in German Silesia (modern day Poland), described as high-sugar content fodder beets (Fischer 1989). The development of sugar beet from fodder beet could explain their presence in the same population after population structure analysis.

Adetunji et al. (2014) studied a sugar beet population composed of elite genotypes and wild accessions. They observed neither the separation between the two subspecies of *B. vulgaris* nor the clustering of the annotated accessions. The fact that they pooled together the elite panel and the wild accessions could have masked the structure within the wild population. They also worked on a relatively small wild population of 91 accessions while we studied a large population.

The use of SNP markers that were discovered in a set of elite lines could have been unsuitable for structure analysis of a worldwide population. The bias brought by the SNPs origin was a hypothesis proposed by Adetunji et al. (2014) to explain the absence of clusters found between the wild beets. However, in our study, the structure observed in the worldwide population was very coherent with databases annotation (species, subspecies, and/or groups), suggesting an interest in using elite SNPs for genetic analyses in exotic populations. Moreover, the structure of the same worldwide

population was also analyzed with 4497 DArTs (diversity arrays technology) markers discovered in this population and the clustering pattern that was obtained with DArTs was similar to the one observed with SNPs (K. Henry, personal communication).

Unlike the worldwide population, the structure of the elite panel was less stable between the clustering methods, and it was really difficult to exhibit a coherent picture. This was probably due to a high level of genetic mixing that regularly combined the subpopulations that could have been created.

The community analysis gave results that were slightly different from the other three clustering methods. Community analysis looks for networks in the graph that are densely connected and are more loosely connected to other parts of the graph. This goal tends to form compact communities, which explains why small communities were found instead of large groups. The choice of the cutoff on the AIS kinship matrix was also an issue when building the graph. We chose a cutoff that seemed reasonable since it was linked to the more significant AIS coefficients, but which was to some extent arbitrary.

Spurious associations between a marker and a trait can be caused in association mapping studies by the existence of a population structure, which is not controlled in the model, causing a seeming LD. Within association panels, the real structure of the material is not totally known. Genetic structure is inferred with various statistical methods using molecular markers. However, the structure model might be mistaken and lead to an under- or overestimation of the true panel structure. The under- and overstructured models may either increase the rate of false positives or false negatives (Mezmouk et al. 2011). Therefore, the choice of the structure model to use for association mapping is quite difficult to make. Mezmouk et al. (2011) showed that the different structure models did not converge to the same number of groups and the impact of the models on the association mapping tests depended on the analyzed traits. In this study, four different approaches of population structure analysis were compared in two very different panels, one composed of elite breeding lines and the other of worldwide accessions. The behavior of the methods to compute genetic structure was different according to the panel. Our results were in accordance with the conclusions of Engelhardt and Stephens (2010). The two most common methods to decipher the population structures (PCA and STRUCTURE application software) gave very coherent results when applying to the worldwide population. Unlike natural evolving populations, artificial elite panels are mixed for breeding purpose and are selected for breeding objectives. This selection process hides the ancestral structure of the elite panels and explains the difficulty of obtaining a coherent picture with different methods of structure inference.

Linkage disequilibrium

For the linkage disequilibrium analysis, we used the usual measure r^2 of LD and those proposed by Mangin et al. (2012) to correct for population structure and relatedness. The new measure r^2_S that corrects for population structure is the partial correlation given a structure matrix S. The measure r^2_V that corrects for relatedness is derived by pre-multiplying the genomic data by a matrix $V^{-1/2}$ (where V is the variance–covariance matrix between individuals), to decorrelate the genotypic observations. If V is not invertible, a Moore–Penrose generalized inverse can be used (Mangin et al. 2012). This V matrix can be, for example, a pedigree-based matrix or a kinship matrix computed with genomic marker data. The two corrections can be made together to obtain the measure r^2_{VS}. We compared two corrections of the structure, the one obtained with k-means on the Patterson's PCA and the admixture proportions provided by STRUCTURE. The LD computed by r^2_S with each of the two above matrices gave comparable results within both the elite panel and the worldwide population. However, the greatest correction was done by r^2_V when using the alike-in-state kinship matrix as V matrix. Moreover, r^2_V and r^2_{VS} gave similar results within the two populations.

The LD as measured by r^2 and modeled by the Hill and Weir's model (1988) extended relatively far, and persisted beyond 10 cM in the elite panel on almost all chromosomes. This was completely different within the worldwide population as it fell below 0.1 before 2 cM for all the chromosomes. As soon as the LD was corrected for the relatedness, it decreased extremely rapidly and was below 0.1 by 1 cM for almost all the chromosomes in both populations, except for the chromosomes 3 and 9 within the elite panel where it persisted a little further. The LD decay curves for the elite panel were just a little superior to those for the worldwide population, showing that the genetic improvement process of the elite panel did not create high LD visible with the marker density that we applied. Adetunji et al. (2014) found that the LD vanished around six cM in their elite panel when they corrected the LD measure by the significant axes of the Patterson's (2006) PCA which was equivalent of using the r^2_S measure. The correction by the r^2_S measure was not sufficient to remove all the bias due to relatedness in elite panels which explains why we found a smaller LD distance when measured with r^2_V. However, in the worldwide population, both corrections were equivalent and our results and those of Adetunji et al. (2014) were similar with a LD that did not persist after 1 cM.

Chromosome 3 exhibited a region of LD in the elite panel, even after correction for structure and relatedness, and this was clearly absent in the worldwide population. These results corroborate those obtained by Adetunji et al. (2014) on a different and smaller elite population.

Rhizomania, caused by beet necrotic yellow vein virus (BNYVV), is perhaps the most important disease of sugar beet because it can cause losses of up to 80 % in sugar yield (McGrann et al. 2009). Nowadays, virtually every commercial variety carries at least one resistance gene located on chromosome 3, *Rz1* (Biancardi et al. 2002; Barzen et al. 1992; Lewellen 1988) from the "Holly" source or *Rz2* from the resistant accession WB42 from Denmark (Scholten et al. 1996, 1999). These genes have provided strong resistance to BNYVV and have been, therefore, intensively selected for. The *Xx* locus involved in restoration of male-fertility in a sterile cytoplasm (CMS), extensively used in breeding, has been also located on chromosome 3 (Schondelmaier and Jung 1997). This system allows crosses to be performed easily by male-sterile plants being placed together with plants that produce pollens, called pollinator. However, the elite panel of the study included only lines belonging to the pollinator pool, in which the *Xx* locus was not under any selection pressure. Due to the constitution of the elite panel, the nuclear restorer locus could not be the cause of a larger LD on chromosome 3 contrary to BNYVV resistance genes, which were under strong selection pressure. The region of high LD in the elite panel was located between 46 and 56 cM on chromosome 3, on our reference map. The resistance gene *Rz1* was located at 51 cM on our reference map, inside the LD block, contrary to the resistance gene *Rz2*, which is located around 20 cM apart from *Rz1* as mentioned by Scholten et al. (1999). Based on these results, we can assume that the LD block found on chromosome 3 corresponds to the *Rz1* gene.

Just as for chromosome 3, LD persisted a little more on chromosome 9 in the elite panel than in the worldwide population. This breeding pattern was more surprising. The white beet cyst nematode (BCN) resistance genes $Hs1^{pro-1}$ (Cai et al. 1997) and *Hs2* (Jäger et al. 2012), and the second bolting locus *B2* (Büttner et al. 2010) are mapped on chromosome 9. Nevertheless, these genes are not widespread in sugar beet breeding germplasm and therefore, they are unlikely to be the cause of the LD region that is observed on chromosome 9.

Chromosomes 1 and 2 were different from other chromosomes because their curves of LD decay obtained within the elite panel could be superimposed on those of the worldwide population. This observation was not surprising for chromosome 1 for which no a priori information was available on presence of important genes. For its part, chromosome 2 is known to carry the early bolting gene *B* (Boudry et al. 1994; Schondelmaier and Jung 1997). The annual habit in *Beta vulgaris* was shown to be controlled by a major dominant gene, commonly referred to as the bolting gene *B*, which promotes the initiation of bolting in long days without prior vernalisation. Because bolting reduces root yield, the occurrence of annual bolting in sugar beets has been strongly selected against during the breeding process. Sugar beet cultivars require vernalisation to bolt and can thus be sown in spring and grown vegetatively until the beets are harvested in the fall. Therefore, we firstly expected a more extended LD in the elite panel than in the worldwide population on chromosome 2. However, nowadays, the selection pressure on this trait in elite material must not be strong since the transition from annual to biennial cultivated beets must have occurred a very long time ago. Recombination must have had plenty of time to remove any extended LD around the *B* locus, contrary to the much more recent introduction of the rhizomania loci on chromosome 3.

The corrected LD measure r^2_{VS} was proved to be linked to the power of the association test in the mixed model of association proposed by Yu et al. (2005). Indeed, there is an analytical formula, based on χ^2 probability distribution, to compute the relation between the power of the detection test at the causal locus and the resulting power at a marker in LD with the causal locus. Indeed, the association test follows asymptotically a χ^2 distribution with 1° of freedom on the null hypothesis and a noncentral χ^2 distribution on the alternative. The noncentrality parameters at a causal locus and at a linked marker are proportional, with a proportionality factor equal to the corrected LD measure r^2_{VS}. This leads, for example, to: if a causal locus has a 90 % chance of detection with a type I error of 10^{-5}, which is possible with a noncentrality parameter of 32.5, a marker in LD with $r^2_{VS} = 0.6$ will be necessary to get a power of 50 % which is not a very high power. A $r^2_{VS} = 0.52$ will still be necessary if the type I error is reduced to 10^{-3}, since the noncentrality parameter at the causal locus is then around 21. We studied the genomic coverage of consecutive markers in LD when LD was superior to a threshold. If we could have a high coverage of the genome for LD superior to 0.6, the power of the association analysis along the genome could surely detect some of the causal loci. With a genomic coverage of barely 1 % with a LD superior to 0.1, the density of the markers and the LD in the worldwide population can be judged as really insufficient to conduct an association analysis. With a genomic coverage between 10 and 35 % depending of the trait heritability with a LD superior of 0.3, the marker density and the LD in the elite panel were more adapted to an association study but they were still limited and could easily lead to a failure to find an associated SNPs. These results seem counter intuitive as the genomic coverage for a low heritability trait was larger than for a high heritability trait, giving the wrong impression that causal locus for low heritability trait would be easier to detect than with higher heritability. However, if a trait has a low heritability, this is because causal loci have small additive effects and are therefore difficult to detect. So even if the genomic coverage is larger the power of

causal loci detection is smaller which resolves the paradox. In conclusion, the r^2 measure gives an optimistic view of what could be expected of a future association study, and the $r^2_{VS}$with a kinship matrix a more pessimistic view; the correct compromise is probably the LD measured by r^2_{VS} with a variance–covariance matrix equal to the identity plus a kinship matrix.

Author contribution statement B.M. and E.G. wrote the manuscript, B.M. and F.S. analyzed the data, K.H. selected and provided the genetic resources, B.D. and P.D. realized the genotyping of the worldwide population and G.W. provided the genomic data of the elite population and genetic mapping information. All authors read and approved the final manuscript.

Acknowledgments This research was carried out with the financial support of the French national research agency (ANR) within the AKER program which is part of the "Programme d'Investissements d'Avenir". AKER is a French research initiative for a sustainable beet improvement with innovative breeding strategies based on allelic variation mining and novel -omic tools. We thank Dr Steve Barnes, SES-VanderHave, for his critical review of the manuscript. We thank two anonymous referees who helped us to greatly improve our discussion.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Adetunji I, Willems G, Tschoep H, Bürkholz A, Barnes S, Boer M, Malosetti M, Horemans S, van Eeuwijk F (2014) Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. Theor Appl Genet 127(3):559–571

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

Biancardi E, Lewellen RT, De Biaggi M, Erichsen AW, Stevanato P (2002) The origin of rhizomania resistance in sugar beet. Euphytica 127:383–397

Biancardi E, Skaracis GN, Steinrücken G, De Biaggi M, Panella L, Lewellen RT, Campbell LG, Yu MH, Stevanato P, McGrath LG (2005) Objectives of sugar beet breeding. In: Biancardi E, Campbell LG, Skaracis GN, De Biaggi M (eds) Genetics and breeding of sugar beet. Science Publishers Inc, Enfield, pp 53–168

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 10:P10008

Bosemark NO (1979) Genetic poverty of the sugar beet in Europe. In: Proceedings of the conference broadening genetic base of crops. Pudoc, Wageningen, the Netherlands, pp. 29–35

Bosemark NO (1989) Prospects for beet breeding and use of genetic resources. In: Report of the international workshop beta genetic resources. International board for plant genetic resources. Rome, Italy, pp. 89–97

Bosemark N (1993) Genetics and breeding. In: Cooke DA, Scott RK (eds) The sugar beet crop. Chapman and Hall, London, pp 67–119

Bosemark NO (2006) Genetics and breeding. In: Draycott AP (ed) Sugar beet. Blackwell Publishing Ltd, Oxford, pp 50–88

Boudry P, Wieber R, Saumitou-Laprade P, Pillen K, van Dijk H, Jung C (1994) Identification of RFLP markers closely linked to the bolting gene B and their significance for the study of the annual habit in beets (Beta vulgaris L.). Theor App Genet 88:852–858

Butterfass T (1964) Die Chloroplastenzahlen in verschiedenartigen Zellen trisomer Zuckerrüben (Beta vulgaris). Z Bot 52:46–77

Büttner B, Abou-Elwafa SE, Zhang W, Jung C, Müller AE (2010) A survey of EMS-induced biennial Beta vulgaris mutants reveals a novel bolting locus which is unlinked to the bolting gene B. Theor Appl Genet 121(6):1117–1131

Cai D, Kleine M, Kifle S, Harloff HJ, Sandal NN, Marcker KA, Klein-Lankhorst RM, Salentijn EM, Lange W, Stiekema WJ, Wyss U, Grundler FM (1997) Positional cloning of a gene for nematode resistance in sugar beet. Sciences 275(5301):832–834

Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks Phys. Rev E 70:066111

Cooke DA, Scott RK (1993) The sugar beet crop. Chapman and Hall Publishers, London, p 675

Csardi G, Nepusz T (2006) The igraph software package for complex network research. Int J Compl Syst 1695, http://igraph.org

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4(2):359–361

Engelhardt BE, Stephens M (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet 6:e1001117

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

Fischer HE (1989) Origin of the 'Weisse Schlesische Rübe' (white Silesian beet) and resynthesis of sugar beet. Euphytica 41:75–80

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

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

Gill NT, Vear KC (1958) Agricultural botany. Duckworth, London

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

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341

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

Jäger S, Hemmrich G, Franke A, Dohm JC, Minoche AE, Himmelbauer H, Capistrano GGG, Harloff HJ, Jung C (2012) Re-sequencing and hybrid assembly strategy of two nematode resistant *Beta vulgaris* translocation lines. In: Plant and animal genome XX, 14–18 Jan 2012, San Diego

Josse J, Husson H, Pagès J (2011) Multiple imputation in PCA. Adv Data Anal Classif 5(3):231–246

Jung C, Herrmann RG, Eibl C, Kleine M (1994) Molecular analysis of a translocation in sugar beet carrying a gene for nematode resistance from *Beta Procumbens*. J Sugar Beet Res 41:27–42

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 39:1151–1155

Kraft T, Hansen M, Nilsson N-O (2000) Linkage disequilibrium and fingerprinting in sugar beet. Theor Appl Genet 101:323–326

Lewellen RT (1988) Selection for resistance to rhizomania in sugar beet. In: Proceedings of the 5th international congress plant pathology. Kyoto, Japan, p. 455

Lewellen RT, Whitney ED (1993) Registration of germplasm lines developed from composite crosses of sugarbeet x Beta maritima. Crop Sci 33:882–883

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

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

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

McDonald JH (2014) Handbook of biological statistics, 3rd edn. Sparky House Publishing, Baltimore

McGrath JM, Saccomani M, Stevanato P, Biancardi E (2007) Beet. In: Kole C (ed) Genome mapping and molecular breeding in plants, vol 5. Springer, Heidelberg, pp 191–207

McGrann GRD, Grimmer M, Mutasa-Göttgens EF, Stevens M (2009) Progress towards the understanding and control of sugar beet rhizomania disease. Mol Plant Pathol 10:129–141

Mezmouk S, Dubreuil P, Bosio M, Décousset L, Charcosset A, Praud S, Mangin B (2011) Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. Theor Appl Genet 122:1149–1160

Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104

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

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. Plos Genet 2:e190

Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Algorithm Appl 10(2):191–218

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69(1):1–14

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

R Core Team (2014) R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. http://www.R-project.org

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76:036106

Rosvall M, Axelsson D, Bergstrom CT (2009) The map equation. Eur Phys J Spec Top 178(1):13–23

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644

Scholten OE, Jansen RC, Keiser LCP, de Block TSM, Lange W (1996) Major genes for resistance to beet necrotic yellow vein virus (BNYVV) in *Beta vulgaris*. Euphytica 91:331–339

Scholten OE, de Block TSM, Klein-Lankhorst R, 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

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

Tracy CA, Widom H (1994) Level-spacing distributions and the Airy kernel. Commun Math Phys 159:151–174

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

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

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