



## Comparison of structure analyses and core collections for the management of walnut genetic resources

Anthony Bernard, Teresa Barreneche, Armel Donkpegan, Fabrice Lheureux,  
Elisabeth Dirlewanger

### ► To cite this version:

Anthony Bernard, Teresa Barreneche, Armel Donkpegan, Fabrice Lheureux, Elisabeth Dirlewanger.  
Comparison of structure analyses and core collections for the management of walnut genetic resources.  
Tree Genetics and Genomes, 2020, 16 (5), pp.76. 10.1007/s11295-020-01469-5 . hal-02948538

**HAL Id: hal-02948538**

**<https://hal.inrae.fr/hal-02948538>**

Submitted on 24 Sep 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



# Comparison of structure analyses and core collections for the management of walnut genetic resources

Anthony Bernard<sup>1,2</sup> · Teresa Barreneche<sup>1</sup> · Armel Donkpegnan<sup>1</sup> · Fabrice Lheureux<sup>2</sup> · Elisabeth Dirlwanger<sup>1</sup> 

Received: 27 July 2020 / Revised: 2 September 2020 / Accepted: 3 September 2020  
© The Author(s) 2020

## Abstract

The preservation of maximum diversity within the smallest number of accessions is one of the challenges of germplasm management, and the assessment of the population structure, the relationships between the accessions, and the construction of core collections are the key steps. The choice of suitable molecular markers is often the starting point. In this study, we analyzed a part of the INRAE walnut germplasm collection, which is highly diverse and unique in Europe, consisting of 150 accessions from American, European, and Asian continents. Based on genotypic data already available, using 13 SSRs or 364,275 SNPs, we showed that the first level of structure is detected equally. We also highlighted empirically that approximately 100 SNPs are needed to obtain similar clustering to 13 SSRs in Principal Coordinate Analysis (PCoA). We constructed eight core collections following two strategies (percentage of total allelic diversity or number of accessions) and two construction methods based on different algorithms (“maximum length subtree” and “entry-to-nearest-entry” methods). We showed that core collections based on few SSR markers are able to capture at least 99.5% of the SNP allelic diversity, irrespective of the construction method used. Then, core collections based on each marker type are highly similar, using both construction methods. All these steps are crucial to identify the suitable tools and methods to improve plant genetic resources management.

**Keywords** Germplasm management · SNP · SSR · Core collection · Walnut

## Introduction

In the context of climate change and human population growth, plant genetic resources (PGR) are of the utmost importance. They constitute the foundations of agricultural sustainability and global food safety (Ulukan 2011; Ogwu et al. 2014), but they are facing conservation issues. Regarding ex situ management of PGR, this mainly involves seeds and some collections are at risk, due in part to the fact that they are generally underfunded and that evaluation and characterization are often

imprecise or inadequate (Fu 2017). In that respect, careful PGR management is crucial from storage to use (Maxted et al. 1997). For clonally propagated perennial species, such as fruit and nut tree species, PGR are generally conserved in ex situ orchards as grafted cultivars, which has its pros and cons: the main advantages are that they can be stored under the climate conditions of their intended use, and can be evaluated during storage; but on the other hand, they require a lot of space and the cost of conservation is significant (Hammer and Teklu 2008).

Nowadays, molecular tools contribute to PGR management (Wambugu et al. 2018), since they can help to find genetically close or synonymous accessions and to create “core collections” which will contain maximum genetic diversity within the smallest number of accessions, leading to a reduction in conservation costs, among other things. Simple Sequence Repeats (SSRs) were frequently used because of their high polymorphism, but the development of high-throughput genomics tools has led to the use of Single-Nucleotide Polymorphisms (SNPs). In the Persian walnut (*Juglans regia* L.), a species which is widely disseminated and grown in many temperate regions, more than 20 publications mention the use of SSRs (Bernard et al. 2018a). This type of markers was

---

Communicated by M. Troggio

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11295-020-01469-5>) contains supplementary material, which is available to authorized users.

✉ Elisabeth Dirlwanger  
elisabeth.dirlwanger@inrae.fr

<sup>1</sup> University Bordeaux, INRAE, Biologie du Fruit et Pathologie, UMR 1332, F-33140 Villenave d'Ornon, France

<sup>2</sup> CTIFL, Centre Opérationnel de Lanxade, 24130 Prignolles, France

particularly used to decipher the genetic diversity and the structure of numerous walnut accessions at a continental scale (Ebrahimi et al. 2016). SSRs were also used to perform a landscape genetic analysis across the Asian range of Persian walnut, species native to the foothills of the Himalayas (Pollegioni et al. 2014). The authors showed that reservoirs of high genetic diversity still exist, confirming the importance of the conservation of walnut genetic resources in this area. Recently, a high-density Axiom™ *J. regia* 700K SNP genotyping array was developed and validated, initiating a novel genomic era in walnut (Marrano et al. 2019a). This genomics tool was used for instance to investigate the genome-wide population structure in Iranian walnut populations (Arab et al. 2019), and to perform genome-wide association studies (Bernard et al. 2020; Marrano et al. 2019b).

As a result, it is fair to wonder how to conduct the tasks related to germplasm management, such as core collections construction, and the choice of the type of marker to be used is one of the first questions. Neutral SSR loci, due to slippage during replication, usually mutate much more frequently than SNP loci, leading to population-specific alleles that are useful for revealing population structure (Tsykun et al. 2017). On the other hand, since they are usually developed in small numbers for one species, they therefore may not reflect genome-wide genetic diversity contrary to SNP loci that are much more frequent in the genome of most species (Ljungqvist et al. 2010). These two types of marker bring different views of the structure, and the merits of each are listed in Guichoux et al. (2011). SSRs and SNPs were compared in order to assess population structure and relatedness in short-lived species such as rice (Singh et al. 2013), maize (Hamblin et al. 2007; Van Inghelandt et al. 2010; Yang et al. 2011), sunflower (Filippi et al. 2015), bean (Müller et al. 2015), and cowpea (Desalegne et al. 2017). These comparisons are rarer for perennials (grapes (Emanuelli et al. 2013) and jujube (Chen et al. 2017)). A recent work in an apple gene bank compared population structure and relatedness using 15 SSRs and more than 15,000 SNPs. However, studies on core collections construction are lacking.

The aim of this paper is to identify the type of marker and the core collections construction methods more relevant for routine walnut germplasm management.

## Materials and methods

### Plant materials and genotypic data

The plant material consists in 150 unique accessions of *Juglans regia* from American, European, and Asian continents. The trees are maintained at the *Prunus* and *Juglans* Genetic Resources Center located in the Fruit Experimental Unit of INRAE in Toulence (latitude 44° 34' 37.442" N–

longitude 0° 16' 51.48" W), near Bordeaux, France (S1 Table). The 150 accessions were selected following two steps. (i) We created a GWAS panel of 170 accessions (Bernard et al. 2020) from SSR and SNP genotyping were described and made publicly available (Bernard et al. 2018b; Bernard et al. 2020).

### Structure analyses

Principal Coordinate Analysis (PCoA) was used to determine the structure patterns among the 150 accessions (Fig. 1). Dissimilarities, based on allelic data, were calculated with 10,000 bootstraps, and transformed into Euclidean distances using a power transformation of 0.5. PCoA was performed using the DARwin 6.0.14 software (Perrier and Jacquemoud-Collet 2006), supplemented by the scatterplot3d R package for 3D visualization.

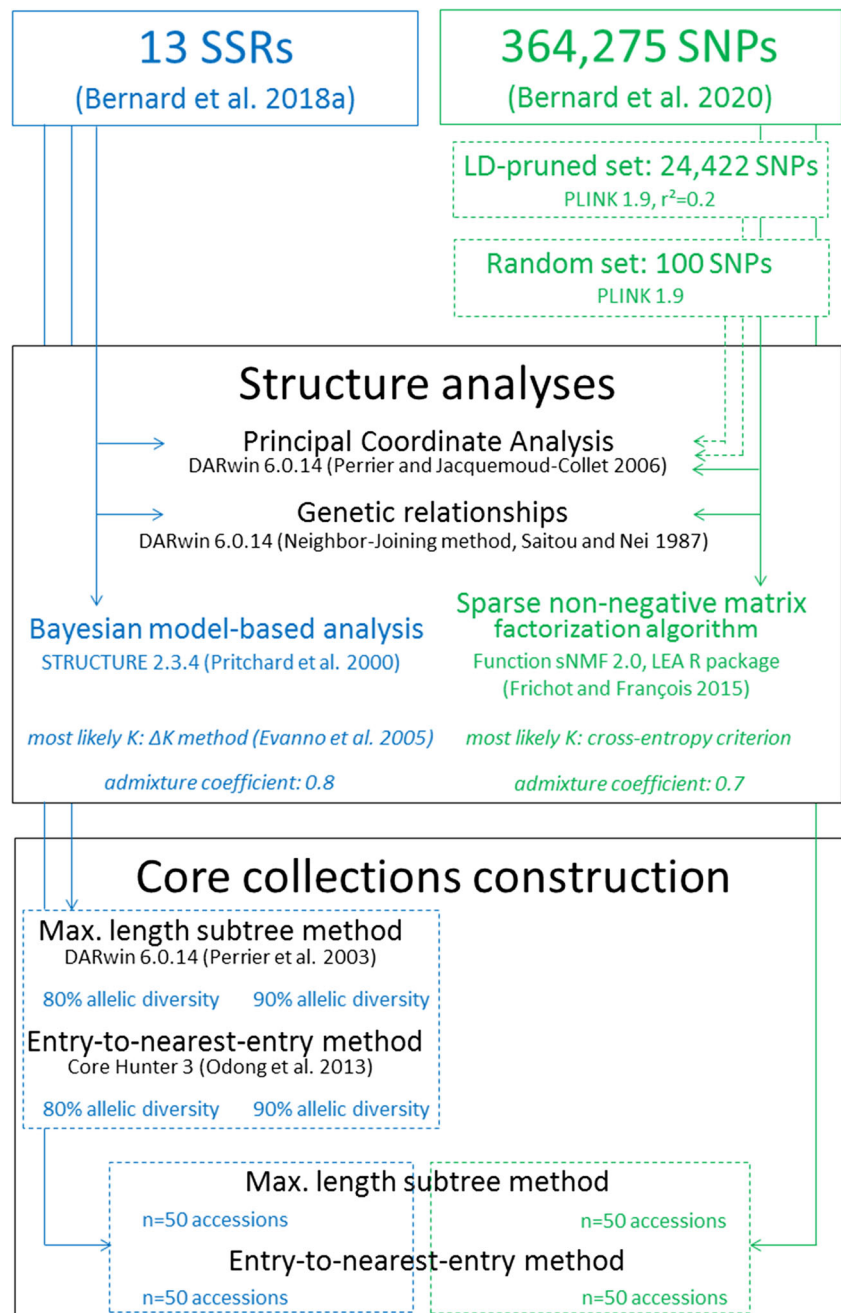
As linked SNPs can account for too much in the population structure variance, particularly in linkage disequilibrium (LD) regions (Price et al. 2008), a pruned subset of SNPs was also used for the PCoA. This filtering was completed using the PLINK 1.9 software, keeping only the SNPs with the higher minor allele frequency, and based on a threshold of  $r^2 = 0.2$  (command= indep-pairwise 50 5 0.2). A subset of 100 SNPs, a number similar to that of SSR alleles, randomly selected, was also tested using the PLINK 1.9 software (command= thin-count 100) to compare PCoA results.

The genetic structure of our panel was also investigated using two types of analyses depending on the markers. Bayesian model-based analysis using the STRUCTURE 2.3.4 software was implemented for the SSR markers (Pritchard et al. 2000), following the method of our previous work (Bernard et al. 2018b). The  $\Delta K$  method (Evanno et al. 2005), implemented in STRUCTURE harvester (Earl and von Holdt 2012), was used to determine the most likely  $K$ .

Using SNP markers, a sparse non-negative matrix factorization algorithm was implemented using the sNMF 2.0 software, available as a function of the LEA R package (Frichot and François 2015). This software has a fast and efficient program for estimating individual admixture coefficients from large genomic datasets, and produces results very close to Bayesian clustering programs such as STRUCTURE (Frichot et al. 2014). The choice of the best number of clusters ( $K$ ) is based on a cross-entropy criterion implemented in the LEA R package. For SSR and SNP markers, thresholds of 0.8 and 0.7 for admixture coefficients, respectively, were chosen to consider one accession as admixed.

Next, the genetic relationships between the 150 accessions were also assessed with the Neighbor-Joining method (Saitou and Nei 1987) using the DARwin 6.0.14 software, for both the SSR and SNP datasets. The Unweighted Neighbor-Joining option was used to build the trees.

**Fig. 1** Workflow of the structure analyses and the core collections construction



### Core collections construction

Several core collections based on both marker types were constructed using two methods and two selection thresholds (Fig. 1). (i) The two methods: we used the “maximum length sub tree” function (Perrier et al. 2003) of the DARwin 6.0.14 software, which looks for a subset of accessions minimizing the redundancy between them, and limiting where possible the loss of diversity (diversity here is expressed by the tree built), and the “entry-to-nearest-entry” method (Odong et al. 2013), implemented in the Core Hunter 3 software (De Beukelaer et al. 2018), which looks for a subset of accessions as different

as possible from each other, avoiding selecting a few clusters of similar accessions at the extreme ends of the distribution. (ii) The two selection thresholds: using the two methods mentioned above, we first used the SSR dataset to capture at least 80% or 90% of total allelic diversity, and the number of alleles captured in one SSR-based core collection was also estimated in the SNP dataset (selection threshold based on percentage of total allelic diversity). Secondly, we used a sampling intensity of 33% ( $n = 50/150$ ) to observe the differences in the selected accessions, using these two methods (selection threshold based on a number of selected accessions).

## Results

### Comparison of the first level of structure using SSR and SNP markers

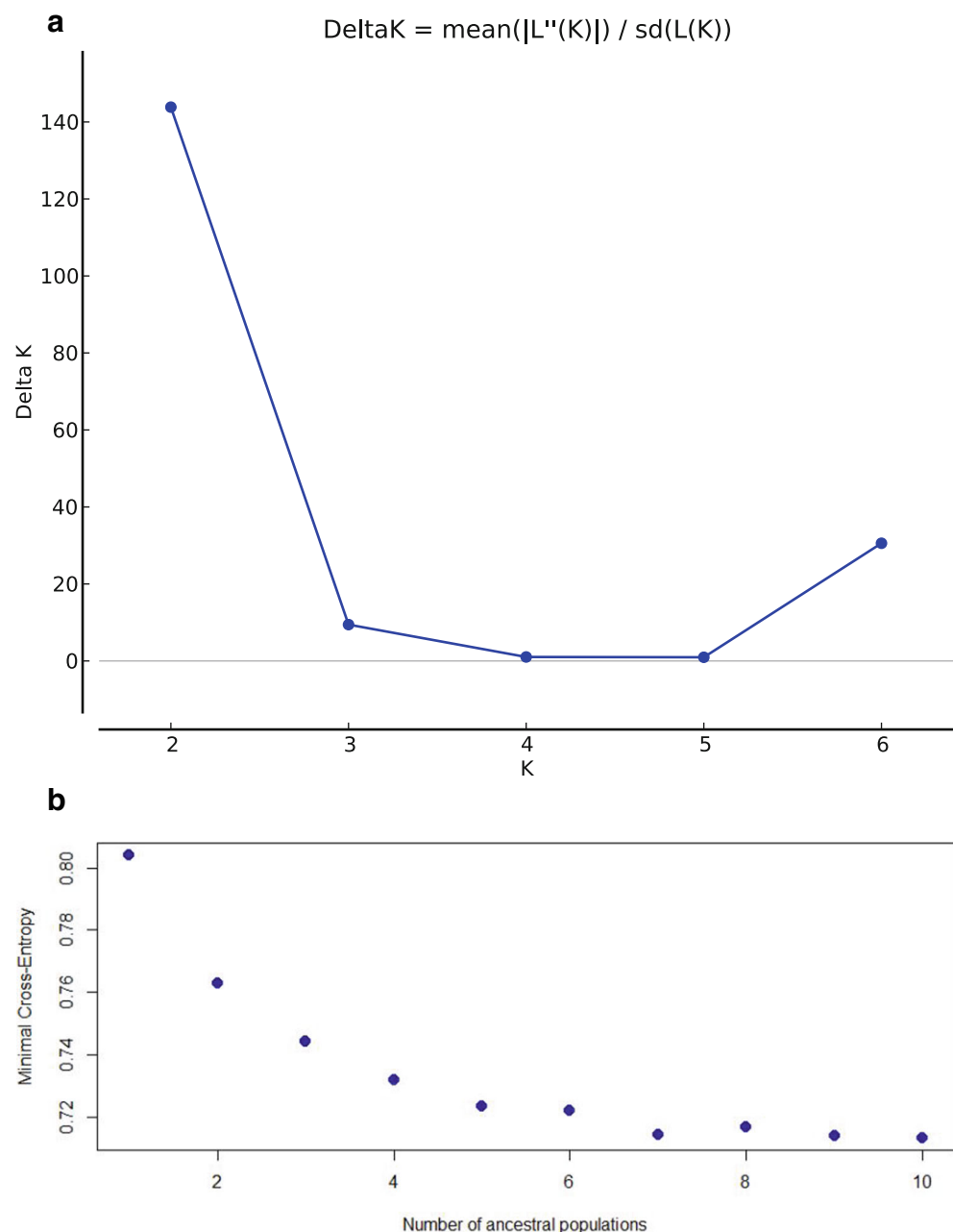
Using SSR markers, the higher drop of  $\Delta K$  is for  $K = 2$ , followed by a rise for  $K = 6$ , and very similar results were obtained using SNP markers, since the higher drop of the cross-entropy criterion is for  $K = 2$ , with a curve slope starting for  $K = 6$  (Fig. 2). To compare the results using SSR and SNP markers, we focused on the first level of structure for  $K = 2$  (S2 Table). The individual admixture coefficients are shown for each of the 150 accessions (Fig. 3). For both marker types, the

clustering is linked to the geographic origin of the accessions, as we already described (Bernard et al. 2018b; Bernard et al. 2020). Group A contains the accessions from Eastern Europe and Asia (named “E”), and group B contains the accessions from Western Europe and America (named “W”).

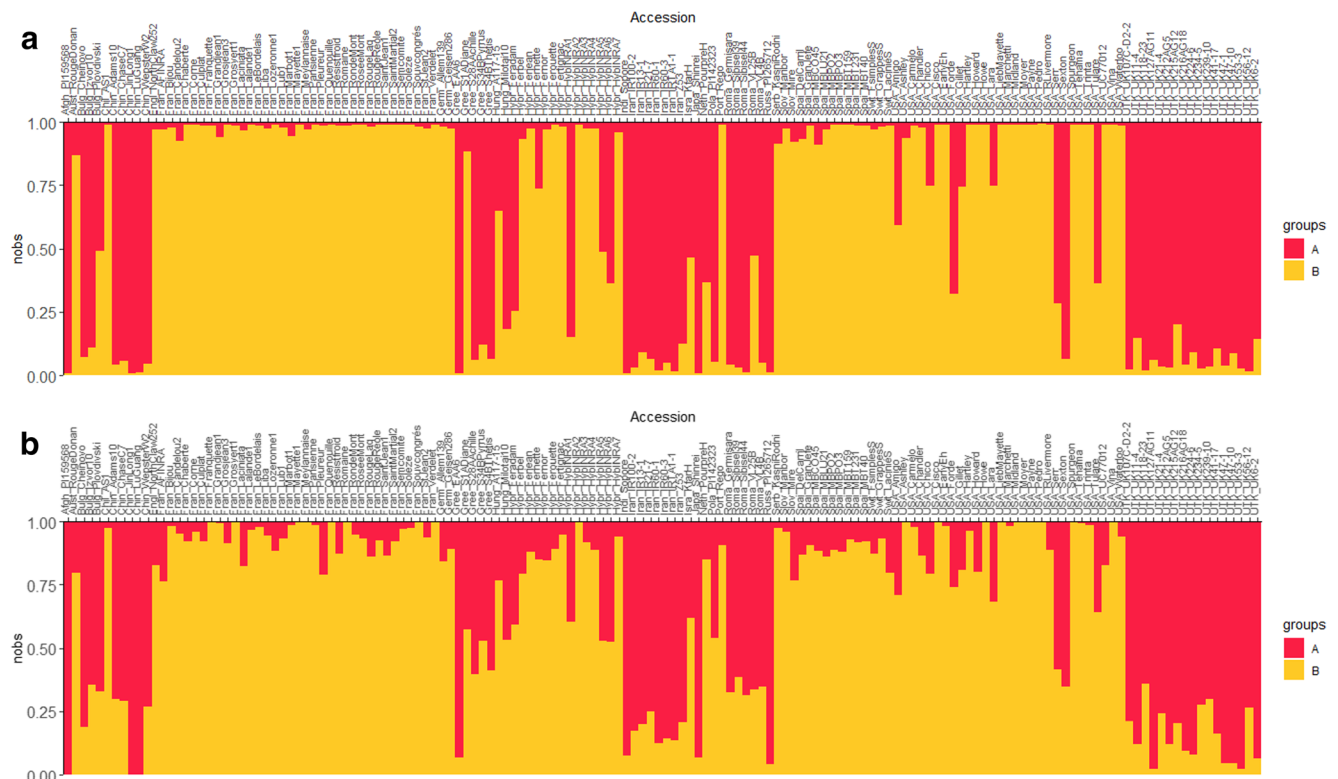
Using SSR markers, with a threshold of 0.8 for the individual admixture coefficient, we found 17 admixed accessions, so 88.7% of the accessions were assigned to a group (Table 1). They include 11 French and Californian modern hybrids and 6 accessions from Eastern Europe, Israel, and Central Asia, confirming our previous SSR analysis.

Using SNP markers and a threshold of 0.7, we found 24 admixed accessions (84% of assignment), including 9 of the

**Fig. 2** The most likely  $K$  subpopulations.  $K$  was evaluated with (a) the  $\Delta K$  method using SSRs, and (b) the cross-entropy criterion using SNPs







**Fig. 3** Bar plots showing the individual admixture coefficients of the 150 accessions for  $K = 2$ . Structure was assessed (a) using SSRs, and (b) using SNPs. The accessions are ordered by their country of origin, in alphabetical order. Group A in red contains the accessions from Eastern

Europe and Asia (“E”), whereas group B in yellow contains the accessions from Western Europe and America (“W”) and hybrids from INRAE

17 found using SSR markers (Fig. 4). From these admixed accessions, 7 are now clustered into group B, and “UK 215AG12” is now in group A. In addition, 14 accessions from group A, and 1 from group B, based on SSR clustering, are found admixed using SNP markers (Fig. 4). Therefore, only 23/150 accessions (15%) are differently clustered. At all events, we did not find any group exchange from A to B, or from B to A, when comparing the clustering based on SSR or on SNP markers.

When using a threshold of 0.8 for the SNPs, the percentage of population assignment for  $K = 2$  is low (70.7%) but is still high for a threshold of 0.7 (84.0%) (Table 1), showing that 0.7 and 0.8 are more appropriate thresholds for SNPs and SSRs, respectively. In addition, we ran a Spearman rank correlation

test for  $K = 2$  and found that the clustering between SSR and SNP markers is highly correlated, up to 84%.

The PCoAs constructed in 2D and 3D show the clustering of the 150 accessions following the results obtained from  $K = 2$  (Fig. 5). For  $K = 2$ , the PCoA results are in agreement with the structure based on SSR or SNP markers, since the scatterplots for groups A and B are well-defined by the first principal component. The admixed accessions are positioned mainly between the two groups. For both methods, the three Manregian walnuts (“Chase C7”, “Wepster W2,” and “Adams 10”) are found to be genetically diverse. Regarding the percentage of explained variation, they are also comparable using both types of markers. The first three axes ( $x$ ,  $y$ , and  $z$ ) explain 21.86% of the cumulative variation for SSRs, and 14.91% for SNPs.

However, we found some differences between the two types of marker. Using SSR markers, the scatterplot corresponding to group A is more extensive, whereas group B is more scattered using SNP markers. Additionally, this group is split in two, showing the French accessions on one side and the Californian ones on the other.

The Neighbor-Joining method implemented in DARwin 6.0.14 was used to construct grouping trees with the 150 accessions (Fig. 6). The main branching groups of the trees

**Table 1** Percentage of population assignment to a group from  $K = 2$  to  $K = 6$ , using SSR and SNP markers

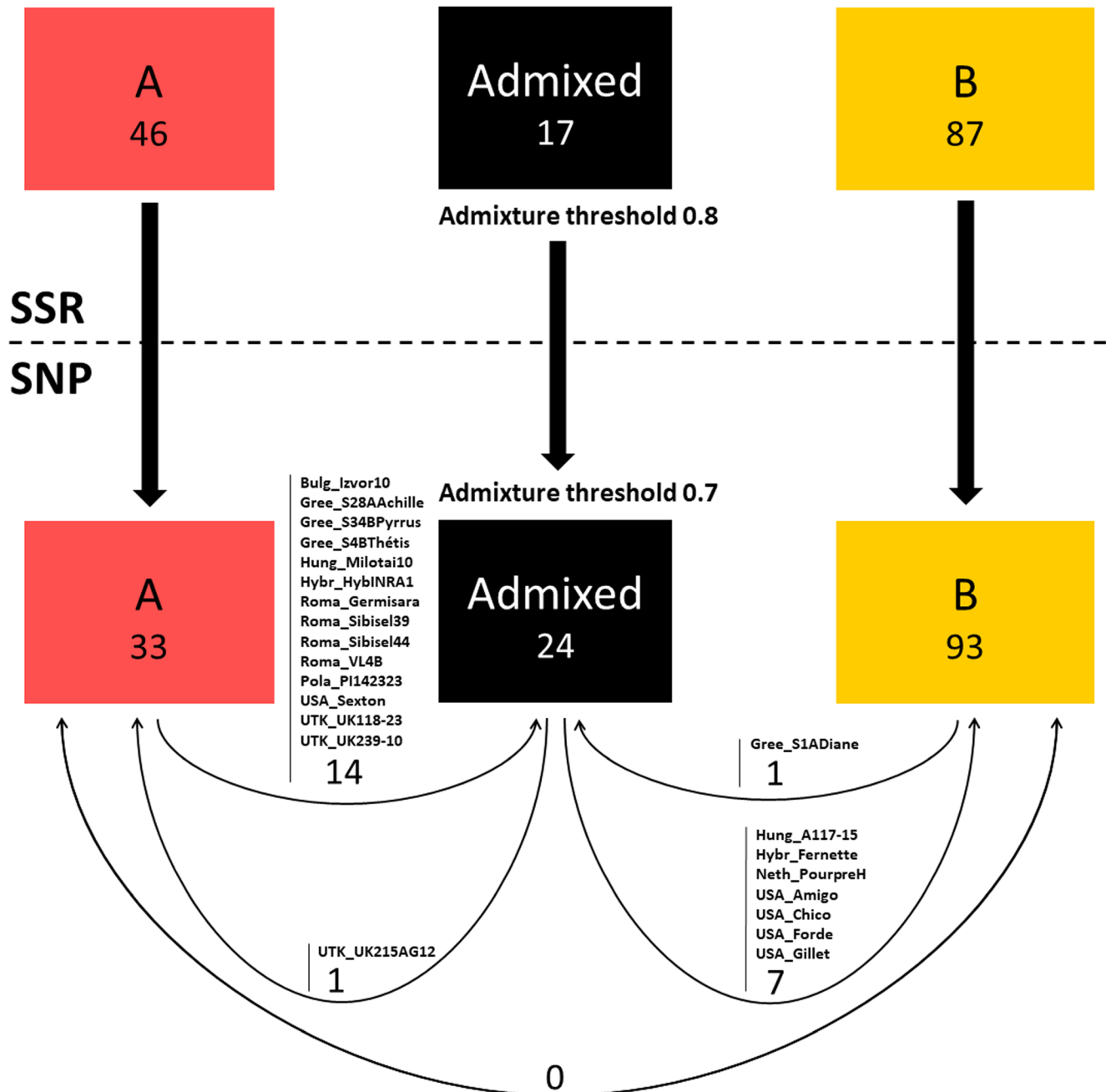
	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
SSR (admixture threshold = 0.8)	88.7	85.3	73.3	59.3	68.7
SNP (admixture threshold = 0.7)	84.0	62.7	56.0	49.3	44.7
SNP (admixture threshold = 0.8)	70.7	47.3	36.0	26.7	22.0

obtained with both markers are in agreement with the structure results ( $K = 2$ ), since they are mainly defined by groups A and B. The two accessions “Jin Long 1” from China and “PI 15 95 68” from Afghanistan are among the longest length branches. However, few differences were detected between the structure and the tree, with accessions from group A found in the branching group mainly corresponding to group B: 9 accessions using SSR markers, and 4 accessions using SNP markers.

We then inferred the individual admixture coefficients of the 150 accessions from  $K = 3$  to  $K = 6$  for both markers (S2 Table). Bar plots are given for information in S1-S4 Figs.

### Comparison of three sets of SNPs for PCoA assessment

In addition, for PCoA assessment, we compared the entire set of 364,275 SNPs with a subset of SNPs filtered for LD, with a



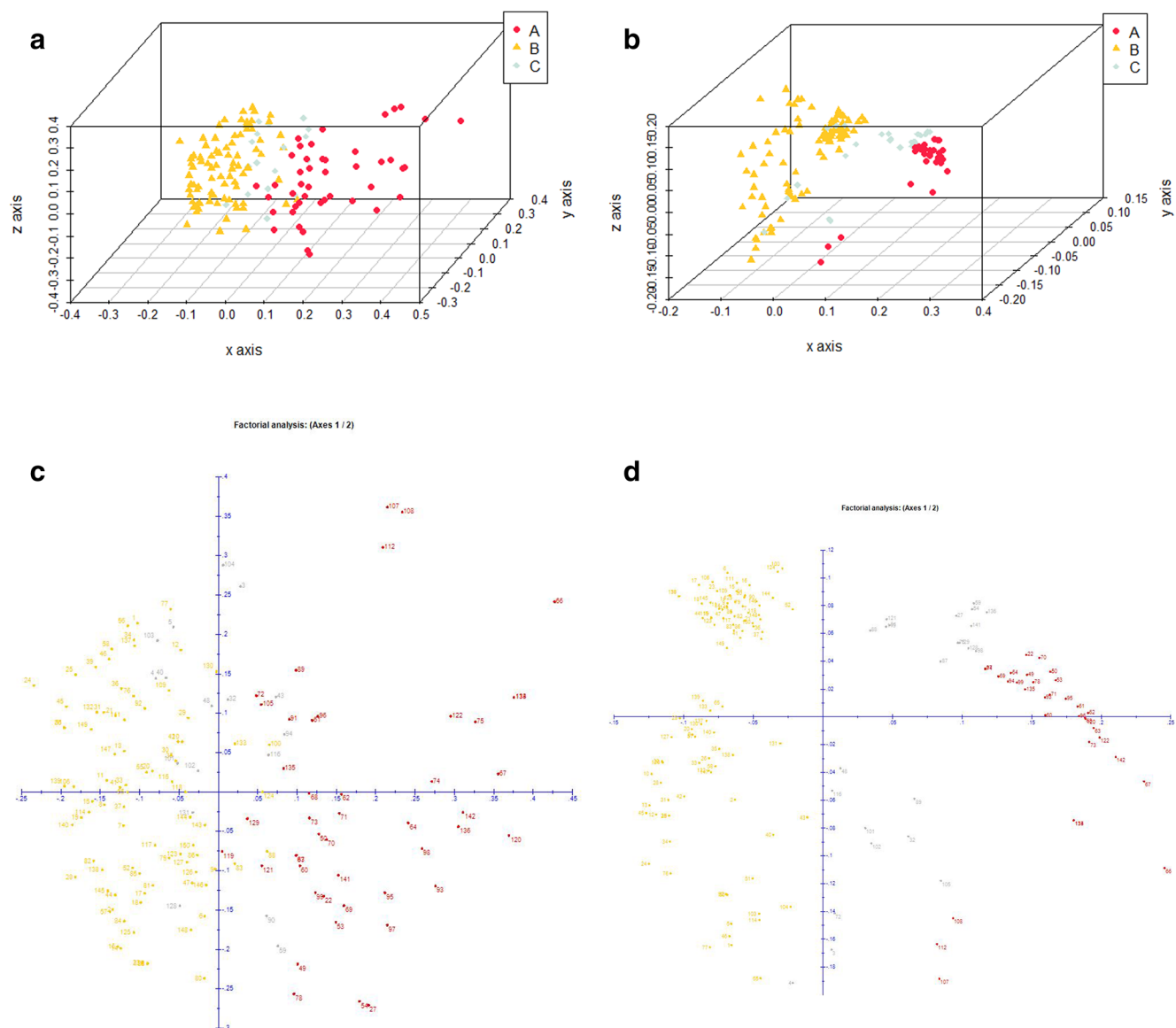
**Fig. 4** SNP-based clustering results, compared with SSR-based clustering results. For a threshold of 0.7 for admixture, SNPs clustered 14 accessions found in group A, and one found in group B, compared with SSRs, into the admixed group. Conversely, SNPs clustered one and seven

accessions found to be admixed, compared with SSRs, into groups A and B, respectively. There was no clustering exchange between groups A and B, when comparing both markers

threshold of  $r^2 = 0.2$  (Fig. 7). We then retained 24,422 SNPs, or 6.7% of the entire set. Interestingly, the results found for both datasets are strongly similar, with the scatterplots well-distinguished, according to the  $K = 2$  results (Fig. 7). We still distinguish the French accessions from the Californian accessions within group B. The main difference is that the variance is better explained by the first three axes using the LD-pruned set (18.90% vs. 14.91%). By comparing the entire set with a random subset of 100 SNPs, a range comparable with the total number of SSR alleles, we found that the scatterplots are less well-defined, but still in agreement with the  $K = 2$  results (Fig. 7). In this case, we cannot distinguish the French accessions from the Californian accessions within group B.

### Comparison of core collections using SSR and SNP markers

In light of the results obtained concerning structure analyses (13 SSRs give similar results compared to 364,275 SNPs), we decided to construct first four core collections based on SSR markers. We created them with both core collections construction methods and two thresholds, to capture at least 80% or 90% of total allelic diversity. To capture at least 80% of total SSR allelic diversity, we need 16 accessions using the method based on the “maximum length sub tree” function of DARwin 6.0.14, and 15 accessions using the “entry-to-nearest-entry” method of Core Hunter 3 (Table 2). Next, to capture at least 90% of allelic diversity, we need 27 and 32 accessions for each construction method, respectively. The accessions



**Fig. 5** Principal Coordinate Analysis scatterplots. PCoAs were constructed in 3D using (a) SSR, and (b) SNP markers, and in 2D using (c) SSR, and (d) SNP markers. The 150 accessions are colored following the  $K = 2$  results: group A in red, group B in yellow, and admixed in grey



belong mainly to group A, from Eastern Europe and Asia, known to be more diverse: 11/16 and 12/15 respectively for 80%, and 20/27 and 21/32 respectively for 90%. They include, for instance, Iranian accessions, the Indian “Sopore” and several accessions from the Botanical Garden of Kiev. Regarding the French accessions, few landraces such as “Corne,” “Marbot” and “Grandjean1” were kept only when capturing 90% of allelic diversity using “Core Hunter 3”.

Moreover, the consistency of the results between both percentages of total captured allelic diversity was checked for the two core collection construction methods. Using Darwin 6.0.14, the 16 accessions selected to capture 80% of allelic diversity were found among the 27 selected to capture 90% (Table 2). However, using Core Hunter 3, 2/15 accessions selected to capture 80% were not found among the 32 selected to capture 90%. Hence, for 80%, 9 accessions (among 16 or 15) are in common between the two methods, and 19 accessions for 90% (among 27 or 32). We then estimated the number of alleles captured in the SNP dataset for each core collection. For the two SSR-based core collections capturing at least 80% of total allelic diversity, they also capture more than 99.5% of total SNP allelic diversity. For 90%, they reach more than 99.9%.

Next, we constructed four additional core collections using the same sampling intensity threshold of 33% to select 50/150 accessions, with both methods and both markers. Results show that 34/50 accessions are in common between the SSR and SNP markers using Darwin 6.0.14, and 32/50 accessions using Core Hunter 3 (Table 3). The accessions belong mainly to group A, from Eastern Europe and Asia, known to be more diverse (between 30 and 31/50 using SSRs, and between 22

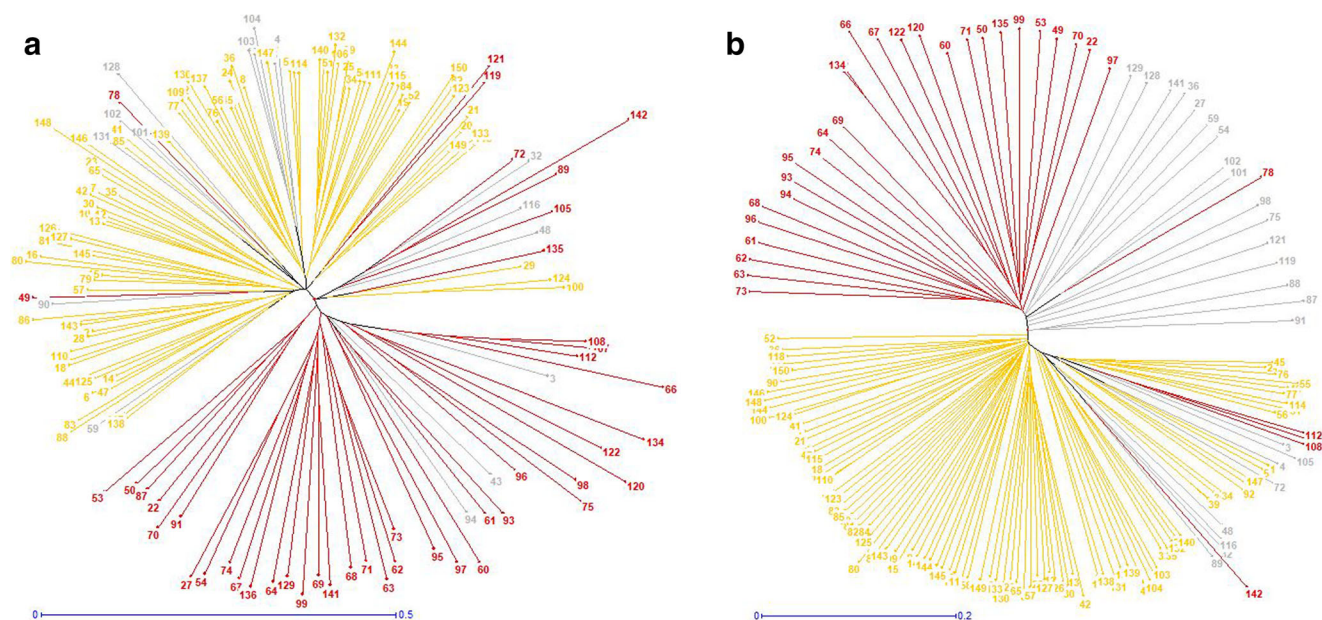
and 23/50 using SNPs). They both include the Iranian accessions, the Chinese “Jin Long 1,” “PI 15 95 68” from Afghanistan, the Indian “Sopore,” the Bulgarian “Izvor 10” and “Plovdivski,” and several accessions from the Botanical Garden of Kiev.

We also checked the consistency of the results between the two core collection construction methods for both markers. Using SSR markers, 39/50 accessions are in common between the two methods, and 45/50 accessions using SNP markers. Regarding the number of alleles, for SSR markers, we captured 97.1% and 94.3% of the 105 total alleles found within the entire collection, using the “maximum length subtree” and the “entry-to-nearest entry” methods, respectively. Using SNP markers, we captured more than 99.9% of the 728,550 total alleles found with the two methods (Table 3).

## Discussion

### The use of either SSR or SNP markers shows comparable results for the collection structure

By genotyping the panel of 150 *J. regia* accessions using 13 SSRs and more than 300,000 SNPs, we obtain similar results. Both types of marker showed a first level of structure for  $K = 2$ , with no exchange of accession between the main groups (A and B), which are related to the geographical origin of the accessions. Similar findings were reported in Ebrahimi et al. (2016), with  $K = 2$  mainly corresponding to a group from Asia, and a group of European and North Africa. The exchanges observed between both types of marker only concern

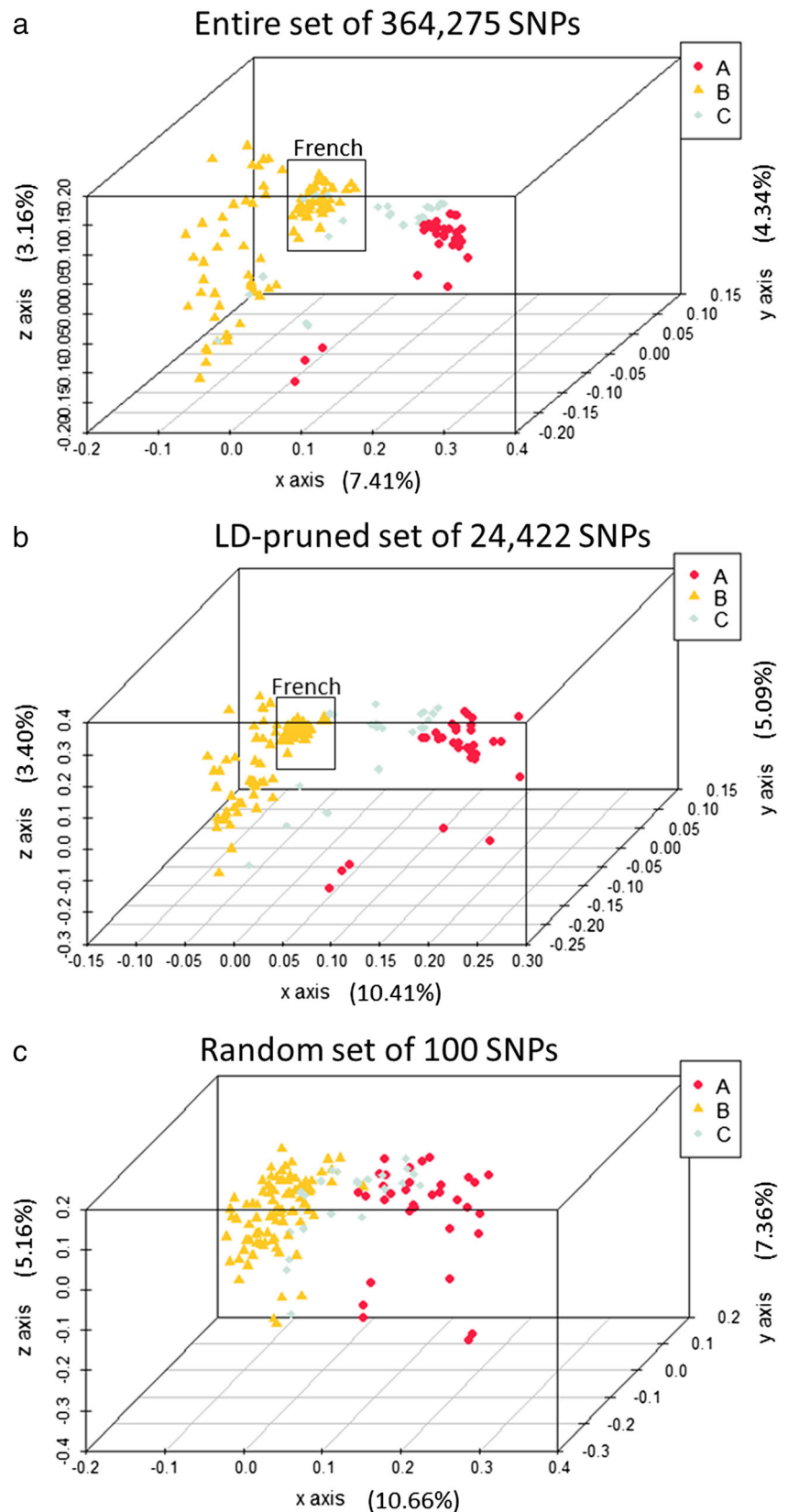


**Fig. 6** Neighbor-Joining trees. Trees were constructed using (a) SSR, and (b) SNP markers. The 150 accessions are colored following the  $K = 2$  results: group A in red, group B in yellow, and admixed in grey

accessions that switched from one group to the admixed cluster, or vice versa. From  $K = 3$  to  $K = 6$ , the results remain

highly comparable, with the highlighting of substructures linked, for instance, to “Payne” pedigree for the Californian

**Fig. 7** Comparison of SNP sets for Principal Coordinate Analysis. PCoA were constructed using (a) the entire set of 364,275 SNPs, (b) the LD-pruned subset of 24,422 SNPs, and (c) the random set of 100 SNPs





Consequently, SNP markers confirmed our results obtained using SSR markers (Bernard et al. 2018b). When considering the LD-pruned set of 6.7% of the entire set of SNPs, PCoA showed consistent clustering patterns and also consistency with those using the 13 SSRs.

These kinds of findings have previously been observed in other works, except that the number of markers needed to obtain a comparable resolution power would be different. For example, the broad patterns of PCoA were similar using 36 SSRs with 2.2 alleles per locus on average and 36 SNPs in 375 Indian accessions of rice (Singh et al. 2013). However, in local accessions of cowpea from East African countries, similar clustering patterns were found using more SNPs than SSRs (151 vs. 13) (Desalegne et al. 2017). In jujube, within a core-collection of 150 accessions, only 18 were classified into different groups based on the results of structure analysis using 24 SSRs and 4680 SNPs (Chen et al. 2017). Within various inbred maize lines, SSRs performed better at clustering accessions into groups using about 10 times more SNPs (Hamblin et al. 2007; Yang et al. 2011). In apple, another clonally propagated perennial, a comparison between 15 SSRs and more than 15,000 SNPs showed strong concordance between both markers for relatedness estimation and that about five to ten SNPs were equal to one SSR (Larsen et al. 2018). Other works suggest the use of three times (Müller et al. 2015) or seven to 11 times more SNPs to obtain comparable informative results (Van Inghelandt et al. 2010). In our study, we took care to choose highly polymorph and robust SSRs suitable for *J. regia* diversity studies, by reviewing the literature (Dangl et al. 2005; Woeste et al. 2002; Dang et al. 2016). When working with biallelic markers such as SNPs, it is known that the genetic distances can be equivalent to those calculated with SSRs using the formula  $n(k^{-1})$ , where  $k$  is the average number of alleles per locus, and  $n$  is the number of loci (Laval et al. 2002). With 13 loci and 8.1 alleles per locus on average (four times that for SNPs), we theoretically need  $[13 \times (8.1^{-1})] = 93$  SNPs to obtain equivalent genetic distances in our panel. Our findings based on PCoA clustering using the random set of 100 SNPs are consistent with this number of SNPs.

Moreover, SNPs globally tend to give higher proportions of inferred admixture, as observed in sunflower (Filippi et al. 2015). Regarding the admixture thresholds chosen, the results found when comparing 0.7 or 0.8 for SNPs also show that 0.7 is clearly more suitable for obtaining comparable structure. In this respect, we found that the percentage of assignment decreases as the  $K$  increases, particularly for the SNPs. Such differences were also reported in maize (Yang et al. 2011) and grape (Emanuelli et al. 2013).

### The choice of marker type will depend on the necessary task related to germplasm conservation or utilization

The management of PGR comprises the following main steps: their conservation, which consists in acquisition of plant material (through the collection or protection of reserves in situ, or by exchange of ex situ material), their maintenance (such as storing and propagation), their characterization (based on both genotype and phenotype), and their utilization for research, breeding programs, or production (Bretting and Widrlechner 1995). Due to the increasing availability of genomics tools, a cross-disciplinary field which aims to use genomics in germplasm management has grown (Jia et al. 2017). But undoubtedly, the choice of SSR or SNP markers will depend on the purposes. For obvious reasons, a first choice criterion could be the cost of genotyping. New genomics technologies have a cost that has decreased dramatically in recent years. Apart from the DNA extraction, the cost of SNP vs. SSR genotyping was about 2600 times less in our study. For guidance only, we paid USD\$8.80 for 13 SSR loci per sample (0.7 USD\$/locus/sample) and USD\$98.90 for one array of 364,275 robust loci per sample (2.7E-4 USD\$/locus/sample). However, SSR genotyping is often more “flexible,” since we could choose precise numbers of loci and samples. In the case of a SNP array, all the loci available are assayed, for a 96-well DNA plate. In addition, only 59.8% of the available SNPs on the array were usable for the analyses after quality control. The cost comparison was also a topic of discussion in Larsen et al. (2018), in which SNP data was 10 times more expensive than SSR data. However, their SSR genotyping required much more manual preparation, such as PCR reactions.

A second choice criterion could be the nature of the plant material managed, particularly if the researcher works only on one crop, or with its wild species. In our case, the SSRs used were also highly transferable into wild species of the genus *Juglans* spp. (Bernard et al. 2018b). Conversely, the Axiom™ *J. regia* 700K SNP array used is valid on the cultivated species *J. regia* only and failed on our few wild species accessions tested. But this is not a general case. In Emanuelli et al.’s work (2013), the set of 384 SNPs used served to analyze 2273 accessions of grapes (*Vitis* spp.), including cultivated grapevines (*V. vinifera* ssp. *sativa*), wild grapevines (*V. vinifera* ssp. *sylvestris*), and non-*Vitis vinifera* species used as rootstocks.

A third reason to use either marker could be the main purpose of the genotyping. Well-chosen neutral SSRs would be sufficient for population structure and relationship determination, particularly in the first steps of germplasm management, since the computational time for analysis is lower. But SNPs, likely to be associated with functional variation, would be preferred for a genome-wide association study purpose (Vargas et al. 2016; Larsen et al. 2018).



## Markers and construction methods must be compatible with the preservation of phenotypic variability

In light of our results, we found that SSR-based core collections could capture more than 99.5% of SNP total allelic diversity. SSR markers, when properly selected using the literature, may be the markers of choice to achieve this goal. Using SSRs, our analyses preferentially selected accessions from East Europe and Asia, as expected, because of their global higher genetic diversity, and helped us to understand that French landraces have a moderate level of genetic diversity. Given that French landraces represent 20% of the entire plant material panel, these findings confirm that their diversity is moderate. However, here it is important to note that the accessions selected in one core collection, even when slightly different, are not always the same in the repeated runs in Core Hunter 3, contrary to in Darwin 6.0.14. This feature, algorithm-dependent and more frequent when the panel has redundancy (Vargas et al. 2016), makes the comparison between construction methods more difficult.

In parallel with the preservation of allelic diversity, it is also necessary to take phenotypic variability into account. The INRAE walnut germplasm collection contains some accessions with unusual traits, such as weeping branches, lacinate leaves, or purple foliage, which may be used for ornamental purposes. Interestingly, the accession with weeping branches “Pleureur” is in the two core collections capturing 90% of SSR total allelic diversity, and the accession with lacinate leaves is in the two core collections constructed using Core Hunter 3. Based on chronological phenotypic data available (Bernard et al. 2019) and new data acquired, we also explored whether the core collections contain a high or low range of variability of some important traits. For the budbreak date for instance, crucial for climate change adaptation, the ten earlier accessions are “Early Ehrhardt,” “Mire,” “Payne,” “Serr,” “Kfar Hanania,” “IR 60-1,” “Sopore,” “Z 53,” “Ashley,” and “Lu Guang,” with a range of budbreak date from 65 to 75 Julian days. Five among the ten are found on the core collection “90% - Darwin.” Conversely, none of the ten later accessions (“Fertignac,” “Le Bordelais,” “St Jean n°1,” “Lalande,” “Candelou,” “Maribor,” “Semence Comité Dordogne,” “Ronde de Montignac,” “Culplat,” and “Romaine”), with a range of budbreak date from 110 to 122 Julian days, is found in the four core collections. Curators are well aware of the importance of phenotype knowledge for an effective PGR management. Currently, new ways to construct core collections are being evaluated to maximize the frequency of minor alleles and phenotypes (Vargas et al. 2016). An interesting application of such “custom methods” in carrot is well-described in Corak et al. (2019). The authors suggest that a core subset based on “custom methods,” e.g., core balancing

functional and genetic diversity based on high-density genomic data and estimated breeding values, would be directly useful for breeders and more efficient for breeding programs.

## Conclusion

In our comparison using 150 *J. regia* accessions, both SSR and SNP markers were highly efficient for highlighting the INRAE walnut collection structure, even though the number of SSRs is much lower than the number of SNPs. Moreover, similar results were obtained for core collections, irrespective of the construction method and the marker type used. The strategy for core collection construction can be based on several criteria and must meet germplasm management requirements (genetic diversity) and constraints (cost and surface of conservation). The lower the number of selected accessions, the higher the probability of discarding a genotype with traits of interest. Thus, this highlights the importance of phenotypic evaluation in germplasm management.

It is therefore important to consider the task of germplasm management when choosing the most appropriate marker. In general, few SSR markers are suitable and sufficient for obtaining a global idea of the structure and the genetic diversity of a germplasm, and for constructing a core collection.

**Acknowledgments** We would like to thank the Fruit Tree Experimental Unit of the INRAE in Toulence and the *Prunus/Juglans* Genetic Resources Center for the maintenance of the collection and for helping us to collect the samples. The CTIFL, as the INNOV’noyer project owner in partnership with INRAE Bordeaux, would like to thank the Région Nouvelle-Aquitaine, and the Cifre convention of ANRT (Agence Nationale de la Recherche et de la Technologie). It is also important to note that the project is supported by the Agri Sud-Ouest Innovation competitiveness cluster. Finally, we would like to thank the late Eric Germain, former head of the breeding program at INRAE Bordeaux from 1977 to 2007. His remarkable work, subsequently pursued by Francis Delort, has given us the opportunity to study a rich set of plant material.

**Code availability** Not applicable.

**Data archiving statement** The SSR genotyping dataset in “tab” format is freely and openly accessed in the Portail Data INRAE repository, via the identifier “INRAE’s Walnut Genotyping Resources” and the following Digital Object Identifier (DOI): <https://doi.org/10.15454/XPKII8>. The dataset called “1\_Diversity\_SSR\_data.tab” is related to a diversity analysis panel of 254 accessions, including the 150 accessions in this study. The additional file called “2\_Diversity\_SSR\_info.txt” gives information concerning the SSR loci.

The SNP genotyping dataset in “hapmap” format is freely and openly accessed via the same DOI. The dataset called “3\_GWAS\_SNP\_hapmap.txt” is related to a GWAS panel of 170 accessions, including the 150 accessions in this study. The additional file called “5\_List\_of\_ID\_SNP.tab” is used to link the array identifier name of the accessions with the identifier name used in this study.

**Authors’ contributions** Anthony Bernard, Teresa Barreneche, and Elisabeth Dirlewanger conceived and supervised the study. Anthony Bernard and Armel Donkpegan implemented the analyses. Anthony

Bernard, Teresa Barreneche, and Elisabeth Dirlewanger interpreted the results. Anthony Bernard wrote the manuscript. Anthony Bernard, Teresa Barreneche, Armel Donkpegan, and Elisabeth Dirlewanger contributed in the discussion of the results. All authors reviewed the manuscript. All authors read and approved the final version of the manuscript.

**Funding** This work has been mainly funded by the “Région Nouvelle-Aquitaine” in the project “INNOV’noyer,” coordinated by the CTIFL, and in partnership with the INRAE of Bordeaux. This work has been also funded by the “Cifre” convention number 2016/1558 of ANRT (Agence Nationale de la Recherche et de la Technologie). INRAE funded part of the PhD grant of Anthony Bernard.

## Compliance with ethical standards

**Competing interests** The authors declare that they have no competing interests.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Arab MM, Marrano A, Abdollahi-Arpanahi R et al (2019) Genome-wide patterns of population structure and association mapping of nut-related traits in Persian walnut populations from Iran using the Axiom *J. regia* 700K SNP array. *Sci Rep* 9:6376
- Bernard A, Lheureux F, Dirlewanger E (2018a) Walnut: past and future of genetic improvement. *Tree Genet Genomes* 14:1
- Bernard A, Barreneche T, Lheureux F, Dirlewanger E (2018b) Analysis of genetic diversity and structure in a worldwide walnut (*Juglans regia* L.) germplasm using SSR markers. *PLoS One* 13(11): e0208021
- Bernard A, Barreneche T, Delmas M, Durand S, Pommier C, Lheureux F, Tranchand E, Naudin M, Dirlewanger E (2019) The walnut genetic resources of INRA: chronological phenotypic data and ontology. *BMC Res Notes* 12:662
- Bernard A, Marrano A, Donkpegan A, Brown PJ, Leslie CA, Neale DB, Lheureux F, Dirlewanger E (2020) Association and linkage mapping to unravel genetic architecture of phenological traits and lateral bearing in Persian walnut (*Juglans regia* L.). *BMC Genomics* 21: 203
- Bretting PK, Widrechner MP (1995) Genetic markers and horticultural germplasm management. *HortScience* 30:1349–1356
- Chen W, Hou L, Zhang Z, Pang X, Li Y (2017) Genetic diversity, population structure, and linkage disequilibrium of a core collection of *Ziziphus jujuba* assessed with genome-wide SNPs developed by genotyping-by-sequencing and SSR markers. *Front Plant Sci* 8:575
- Corak KE, Ellison SL, Simon PW, Spooner DM, Dawson JC (2019) Comparison of representative and custom methods of generating core subsets of a carrot germplasm collection. *Crop Sci* 59:1107–1121
- Dang M, Zhang T, Hu Y, Zhou H, Woeste K, Zhao P (2016) *De novo* assembly and characterization of bud, leaf and flowers transcriptome from *Juglans regia* L. for the identification and characterization of new EST-SSRs. *Forests* 7:247–263
- Dangl GS, Woeste K, Aradhya MK, Koehmstedt A, Simon C, Potter D, Leslie CA, McGranahan G (2005) Characterization of 14 microsatellite markers for genetic analysis and cultivar identification of walnut. *J Am Soc Hortic Sci* 130:348–354
- De Beukelaer H, Davenport GF, Fack V (2018) Core Hunter 3: flexible core subset selection. *BMC Bioinformatics* 19:203
- Desalegne BA, Dagne K, Melaku G, Ousmane B, Fatokun CA (2017) Efficiency of SNP and SSR-based analysis of genetic diversity, population structure, and relationships among cowpea (*Vigna unguiculata* (L.) Walp.) germplasm from East Africa and IITA inbred lines. *J Crop Sci Biotechnol* 20:107–128
- Earl DA, von Holdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* 4:359–361
- Ebrahimi A, Zarei A, Lawson S, Woeste KE, Smulders MJM (2016) Genetic diversity and genetic structure of Persian walnut (*Juglans regia*) accessions from 14 European, African, and Asian countries using SSR markers. *Tree Genet Genomes* 12:114
- Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troggio M, Myles S, Martinez-Zapater JM, Zyprian E, Moreira FM, Grando MS (2013) Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biol* 13:39
- 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
- Filippi CV, Aguirre N, Rivas JG, Zubrzycki J, Puebla A, Cordes D, Moreno MV, Fusari CM, Alvarez D, Heinz RA, Hopp HE, Paniego NB, Lia VV (2015) Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. *BMC Plant Biol* 15:52
- Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. *Methods Ecol Evol* 6:925–929
- Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196:973–983
- Fu YB (2017) The vulnerability of plant genetic resources conserved ex situ. *Crop Sci* 57:2314–2328
- Guichoux E, Lagache L, Wagner S et al (2011) Current trends in microsatellite genotyping. *Mol Ecol Resour* 11:591–611
- Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One* 2(12): e1367
- Hammer K, Teklu Y (2008) Plant genetic resources: selected issues from genetic erosion to genetic engineering. *J Agr Rural Dev Trop* 109: 15–50
- Jia J, Li H, Zhang X, Li Z, Qiu L (2017) Genomics-based plant germplasm research (GPGR). *Crop J* 5:166–174
- Laval G, San Cristobal M, Chevalet C (2002) Measuring genetic distances between breeds: use of some distances in various short term evolution models. *Genet Sel Evol* 34:481–507



- Larsen B, Gardner K, Pedersen C, Ørgaard M, Migicovsky Z, Myles S, Toldam-Andersen TB (2018) Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PLoS One* 13(8):e0201889
- Ljungqvist M, Åkesson M, Hansson B (2010) Do microsatellites reflect genome-wide genetic diversity in natural populations? A comment on. *Mol Ecol* 19:851–855
- Marrano A, Martínez-García PJ, Bianco L, Sideli GM, di Pierro EA, Leslie CA, Stevens KA, Crepeau MW, Troggio M, Langley CH, Neale DB (2019a) A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density Axiom™ *J. regia* 700 K SNP genotyping array. *Plant Biotechnol J* 17: 1027–1036
- Marrano A, Sideli GM, Leslie CA, Cheng H, Neale DB (2019b) Deciphering of the genetic control of phenology, yield, and pellicle color in Persian walnut (*Juglans regia* L.). *Front. Plant Sci* 10:1140
- Maxted N, Ford-Lloyd BV, Hawkes JG (1997) Complementary conservation strategies. In: Maxted N, Ford-Lloyd BV, Hawkes JG (eds) *Plant genetic resources: the in situ approach*. Chapman and Hall, London, pp 15–40
- Müller BSF, Pappas GJ, Valdisser PAMR et al (2015) An operational SNP panel integrated to SSR marker for the assessment of genetic diversity and population structure of the common bean. *Plant Mol Biol Report* 33:1697–1711
- Odong T, Jansen J, Van Eeuwijk F, van Hintum TJ (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theor Appl Genet* 126:289–305
- Ogwu MC, Osawaru ME, Ahana CM (2014) Challenges in conserving and utilizing plant genetic resources (PGR). *Int J Genet Mol Biol* 6: 16–22
- Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC (eds) *Genetic diversity of cultivated tropical plants*. Science Publishers, Enfield, pp 43–76
- Perrier X, Jacquemoud-Collet J (2006) DARwin software. Available from: <http://darwin.cirad.fr/>. Accessed 1 Nov 2019
- Pollegioni P, Woeste KE, Chiocchini F, Olimpieri I, Tortolano V, Clark J, Hemery GE, Mapelli S, Malvolti ME (2014) Landscape genetics of Persian walnut (*Juglans regia* L.) across its Asian range. *Tree Genet Genomes* 10:1027–1043
- Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD, Goldstein DB, Reich D (2008) Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 83:132–135
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi RK, Singh NK, Singh R (2013) Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS One* 8(12):e84136
- Tsykun T, Rellstab C, Dutech C, Sipos G, Prospero S (2017) Comparative assessment of SSR and SNP markers for inferring the population genetic structure of the common fungus *Armillaria cepistipes*. *Heredity* 119:371–380
- Ulukan H (2011) The use of plant genetic resources and biodiversity in classical plant breeding. *Acta Agric Scand Sect B* 61:97–104
- Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theor Appl Genet* 120:1289–1299
- Vargas AM, de Andrés MT, Ibáñez J (2016) Maximization of minority classes in core collections designed for association studies. *Tree Genet Genomes* 12:28
- Wambugu PW, Ndjondjop MN, Henry RJ (2018) Role of genomics in promoting the utilization of plant genetic resources in genebanks. *Brief Funct Genomics* 17:198–206
- Woeste K, Burns R, Rhodes O, Michler C (2002) Thirty polymorphic nuclear microsatellite loci from black walnut. *J Hered* 93:58–60
- Yang X, Xu Y, Shah T, Li H, Han Z, Li J, Yan J (2011) Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. *Genetica* 139:1045–1054

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.