agronomy

Article

Analysis of Genetic Diversity and Development of A Core Collection in Walnut (*Juglans regia* L.) Germplasm Native to China via Genotyping-by-Sequencing

Jing Ren [1], Yu-An Wang [1], Xiao-Kang Zhou [2], Kai-Wen Xie [2], Fu-Jun Han [1], Hai Peng [1] and Xiao-Yong Liu [1,*]

[1] Institute of Fruit and Floriculture, Gansu Academy of Agricultural Sciences, Lanzhou 730070, China; mailrenjing@163.com (J.R.); wya30@163.com (Y.-A.W.); hanfujun2007@sina.com (F.-J.H.); phai2023@163.com (H.P.)

[2] Tianshui Pomology Institute, Tianshui 741000, China; zhxk_79@163.com (X.-K.Z.); xiekaiwen2023@163.com (K.-W.X.)

* Correspondence: xyliu1966@163.com; Tel.: +0931-7611733

Abstract: Popular knowledge of the population structure and genetic diversity of a plant species is essential for designing improvement strategies. The genotyping-by-sequencing (GBS) approach has been used to simplify complex genomes and has become a popular high-throughput molecular tool for selecting and breeding many crop plants, including those with large genomes. This study aimed to construct a core collection of walnut (*Juglans regia*) germplasm using the GBS approach. A diversity panel of 87 walnut initial genotypes, including 25 landraces, 12 cultivars, and 50 seedling populations, mostly native to the Gansu Province of China, was subjected to GBS. A total of 110,497 high-quality SNPs were identified and used for determining distinct clusters and an optimum number of subpopulations. Structure analysis divided the genotypes into three distinct groups, which coincided with their collection site and year, suggesting a certain degree of separation in the geographical origin and pedigree among the three groups. To maximize germplasm utilization, the genotypes were posteriorly grouped according to the subgroups obtained through GBS analysis. To minimize subsample redundancy, the core collection was designed using a set of 6540 SNPs distributed across all 16 chromosomes. Finally, a core collection comprising nine walnut genotypes (10% of the entire genotype set), including five cultivars, three seedling populations, and one landrace, was assembled. Genetic structure analysis indicated that the core collection has an uneven distribution in the landrace collection, which could be related to environmental conditions, and the genotypes of the landrace collection are similar. Overall, the results of this study and the establishment of the core collection will facilitate the improvement of walnut in future breeding programs.

Keywords: walnut; genotyping-by-sequencing; core collection; genetic diversity

Citation: Ren, J.; Wang, Y.-A.; Zhou, X.-K.; Xie, K.-W.; Han, F.-J.; Peng, H.; Liu, X.-Y. Analysis of Genetic Diversity and Development of A Core Collection in Walnut (*Juglans regia* L.) Germplasm Native to China via Genotyping-by-Sequencing. *Agronomy* **2023**, *13*, 1786. https://doi.org/10.3390/agronomy13071786

Academic Editor: Petr Smýkal

Received: 16 May 2023
Revised: 28 June 2023
Accepted: 29 June 2023
Published: 2 July 2023

1. Introduction

The common walnut (*Juglans regia* L.; Juglandaceae) is a famous cultivated nut with high economic and ecological value [1]. The consumption of *J. regia* by humans can be traced back to Persia (7000 BCE) [2]. As an important nut crop, *J. regia* is cultivated in temperate regions across the world. Today, China is the leading producer of walnuts, accounting for 50% of the annual production worldwide [3]. Previous investigations indicate that the evolutionary history of the common walnut first occurred in the highland region and spread eastward into China via human migration along the Silk Road [4]. There, gene flow became a significant factor influencing the genetic structure of common walnut populations, followed by climate change and geographical transition [5–8]. The genus Juglans contains more than 20 diploid (2n = 2x = 32) species [9]. Among these species, *J. regia*, *J. mandshurica* (Manchurian walnut), *J. sigillata* (Iron walnut), *J. hopeiensis* (Ma walnut), and *J. cathayensis* (Chinese walnut) are widely distributed in China [10,11]. Given the continued uncertainty regarding the location and number of refugia for *J. regia*, the geographic range of its origin remains unclear; however, it is possible that *J. regia* originated from refugia in China [12,13].

Most of the walnut species are monoecious [14]. High genetic variation in the natural populations of walnut may be attributed to sexual reproduction [15]. The existence of protandry, which often leads to outcrossing, increases genetic variation within a species and affects population structure. Under these circumstances, sexual reproduction further enhances segregation among walnut populations, leading to high genetic diversity. These genetic populations can serve as foundational germplasm in a variety of improvement programs. Sexually reproducing genotypes have been used for breeding cultivars with more effective nutrient absorption and utilization capacities and have led to the identification of candidate genes responsive to stress conditions such as extreme temperature, drought, and high salinity/alkalinity. Walnut genotypes capable of sexual reproduction are usually more suitable for the local environment than introduced varieties. With continued cycles of cultivation, harvesting, and selection, walnut genotypes retained by farmers adapt to the local agroecosystem and environmental conditions, similar to how the ecotypes of wild species adapt to the local environment over time for survival [16]. However, a comprehensive and systematic understanding of walnut germplasm resources is lacking, and a number of different genotypes with a common name (homonyms) and identical genotypes with different names (synonyms) have been developed in walnut, affecting the commercialization of walnut products. Moreover, ancient walnut germplasm resources, which are well-adapted to the local environment and contain valuable genes, are constantly decreasing. Therefore, the collection, preservation, and evaluation of local *J. regia* germplasm resources are crucial for its genetic improvement [17]. Thus far, molecular marker methods based on DNA have proven to be stable and reliable for genetic studies of walnut populations [18]. Although, these molecular marker methods have largely solved the identification and screening of walnut genotypes, the massive size of germplasm resources makes the selection process for breeding targets extremely tough, which limits the subsequent utilization of these resources in research studies. To overcome this problem, researchers developed the concept of a core collection [19], which is a subset of the existing germplasm resources that represents the genetic variability of the whole germplasm collection. The purpose of a core collection is to simplify the management of a large number of genotypes while maintaining as much genetic diversity as possible. Core collections can be established based on different types of data, such as ecogeographical data, phenotypic data, and agronomic traits [20–22].

Recently, with advances in molecular marker and next-generation sequencing (NGS) technologies, the genotyping-by-sequencing (GBS) approach has become increasingly popular for large-scale germplasm characterization, which facilitates the identification of high-performance genotypes and accelerates the development of highly efficient breeding strategies by enabling the selection of chromosome-wide genetic diversity [23]. The GBS method is highly suitable for the effective management of germplasm resources. Core collections of some fruit tree species have been created to date [24–26]. Previously, studies have been conducted to investigate the population structure of common walnut, and to establish its core collection based on simple sequence repeat (SSR) [16,27], single nucleotide polymorphism (SNP) [28], amplified fragment length polymorphism (AFLP) [29], and random amplified polymorphic DNA (RAPD) markers [30].

Humans traded walnuts along the ancient Silk Road, scattering common walnut genes across huge natural barriers and great geographical distances [5]. Gansu is an important province in China on the northern Silk Road. The gene exchange between native and Persian walnut germplasm resources obtained a large number of new walnut genotypes; however, these germplasm resources are not well characterized. Therefore, in this study, we aimed to evaluate the genetic diversity and structure of local walnut populations, establish a core collection of common walnuts using the GBS technology, and consequently develop large numbers of SNP datasets. The molecular data reported in this study, for the first time, provide a detailed inventory of the local walnut genetic resources in Gansu Province, China.

2. Materials and Methods

2.1. Plant Material

A set of 87 walnut genotypes (Supplementary Table S1) grown at the Walnut Germplasm Resources Plot of Gansu Academy of Agricultural Sciences (106°.13′ E, 34°.749′ N, Figure 1) were used in this study, including 50 seedling populations (specific origins unclear), which have been used mainly for the selection of superior plants and exploration of germplasm resources; 25 representative landraces, 2 of which represent ancient (800 to 1000-year-old) walnut genetic resources; and 12 introduced varieties registered mainly in China, Japan, and USA. These resources are suitable for cultivation in the Gansu Province of China. Fresh, healthy leaves were collected from each genotype, frozen in liquid nitrogen, and stored at −80 °C.

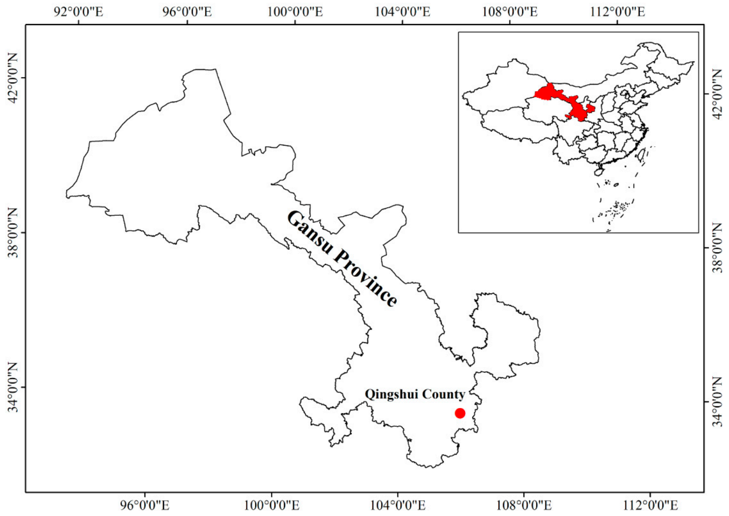

Figure 1. Location of sampling site.

2.2. DNA Isolation

Genomic DNA was extracted from the frozen leaves of all 87 genotypes using the Plant Genomic DNA Kit (TIANGEN, Beijing, China), according to the manufacturer's instructions. DNA degradation and contamination were monitored on 1% agarose gels. DNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, Westlake Village, CA, USA). DNA concentration was measured using the Qubit® DNA Assay Kit with Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA).

2.3. Library Preparation

GBS is an efficient method used for large-scale genotyping based on a reduced representation library (RRL) and high-throughput sequencing. First, a GBS pre-design experiment was performed. Restriction enzymes and sizes of the digested DNA fragments were evaluated using training data. Three criteria were considered to improve the efficiency of GBS: (i) suitable number of tags; (ii) even distribution of enzymatic tags; and (iii) no repeat tags. To maintain the sequence depth uniformity of different fragments, a tight length range was selected (approximately 50 bp).

Next, the GBS library was constructed using the pre-designed scheme. Genomic DNA was incubated at 37 °C with MseI (New England Biolabs [NEB]), T4 DNA ligase (NEB), ATP (NEB), and MseI Y-adapter N-containing barcode. Restriction-ligation reactions were heat-inactivated at 65 °C, and then digested with NlaIII (NEB) and EcoRI (NEB) at 37 °C. The restriction ligation samples were purified with Agencourt AMPure XP (Beckman). Then, PCR was performed using the purified samples, Phusion Master Mix (NEB), universal primers, and index primers with complete i5 and i7 barcodes. The PCR products were purified using Agencourt AMPure XP (Beckman), pooled, and separated by electrophoresis on a 2% agarose gel. Fragments 350–400 bp in size (including indexes and adaptors) were isolated from the gel using the Gel Extraction Kit (Qiagen, Hilden, Germany), purified using Agencourt AMPure XP (Beckman, Brea, CA, USA), and diluted for sequencing.

2.4. Illumina Sequencing

The purified PCR products were sequenced on an Illumina high-throughput sequencing platform to generate 150-bp paired-end reads. SNP genotyping and evaluation were then performed. Sequence reads obtained from each sample were sorted according to the barcodes. To ensure that the reads are reliable and without artificial bias (caused by low-quality paired-end reads, which mainly resulted from duplicate base-calling and adapter contamination), raw data in fastq format were passed through a series of quality control (QC) procedures using in-house C scripts to remove reads containing $\geq$10% unidentified nucleotides (Ns), >50% low-quality bases (Phred quality score [Q] < 5), and adapter sequences (>10 nt aligned to the adapter, allowing $\leq$10% mismatches).

2.5. Mapping to the Reference Genome

Burrows-Wheeler Aligner (BWA) [31] was used to align the clean reads of each sample against the reference genome (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/411/555/GCF_001411555.2_Walnut_2.0/GCF_001411555.2_Walnut_2.0_genomic.fna.gz, accessed on 28 June 2023). Sequence alignment and merging resulted, which were converted to bam files using the SAMtools program [32]. If multiple paired-end reads showed identical external coordinates, only one paired-end read showing the highest mapping rate was retained.

2.6. SNP Detection and Annotation

Variant calling was performed for all samples based on the reference genome sequence using the SAMtools program [32]. SNPs with <20% missing data and <0.05 minor allele frequency (MAF) were used for subsequent analysis.

2.7. Genetic Diversity and Population Structure Analyses

Phylogenetic analysis was conducted in MEGA 6 [33], and the evolutionary history was inferred using the neighbor-joining method. The phylogenetic tree was drawn to scale, and genetic distances were calculated using the number of differences method and expressed as the number of base-pair differences per sequence. To ensure the reliability of genetic relationships among branches, a total of 1000 bootstrap replicates were performed. All positions containing gaps, missing data, and heterozygous sites were removed. The phylogenetic tree was optimized using TreeView and NJplot version 2.3 [34].

To further understand the patterns of genetic relatedness, principal component analysis (PCA) was carried out using the 'prcomp' function in R [35]. Population structure was examined using STRUCTURE version 2.3.4 [36] to investigate the hypothetical number of subpopulations (K) and to determine the ancestry partitioned to each genotype from the inferred subpopulations. The ADMIXTURE method [37] was executed with correlated allele frequencies and burn-in length of 100,000 iterations, followed by 100,000 Markov chain Monte Carlo iterations. The following parameter setting was defaulted to the manufacturer's recommended values. To identify the best K value, cross-validation error was tested for K varying from 2 to 15. After considering 10-fold cross-validations, K with the lowest cross-validation error was selected as the best K value.

2.8. Core Collection Selection

To maximize germplasm utilization, a core subset was selected based on individual genetic diversity representative of the common genetic variation present within the collection. The R package for Core Hunter version 3.2.1 [38] was used to generate the core collection assembly, and 10 replications were performed to achieve repeated selection. Genotypes captured by Core Hunter were compared with the phylogenetic tree to evaluate the core subset's representation of the collection's overall structure.

3. Results

3.1. Summary Statistics of SNP Calling

Statistics of the sequence data generated from the 87 genotypes after eight Ion Proton runs are summarized in Supplemental Table S2. A total of 31,980,496,896 raw sequence reads were obtained from all genotypes. After removing low-quality reads, adapter sequences, and unique alignments, 31,980,287,808 (99%) clean reads were obtained, with an average GC content of 38.26%. In the clean reads, 96.06% and 89.56% of the base calls showed average Q values of >20 (Q20) and >30 (Q30), respectively. Only SNPs scoring Q20 were used for the assessment of genetic variation. The average coverage of the genotypes was 97.26%, with at least 1× coverage site in the reference genome at 11.91%, and the average sequencing depth was 8.89× (Table 1). Using SAMtools, a total of 110,497 (6.84%) SNPs were aligned to the reference genome based on two filtering criteria (missing data < 20%, MAF > 5%), and were finally selected for genetic diversity analysis.

Table 1. Statistics of sequencing depth and coverage.

Statistic	Statistical
Mapped reads [a]	2,482,575
Mapping rate [b]	97.26%
Average sequencing depth [c]	8.89×
Coverage 1× [d]	11.91%
Coverage 4× [e]	5.57%

[a] Number of clean reads compared to the reference genome. [b] The percentage of clean reads that mapped to the reference genome. [c] Average sequencing depth. [d] The percentage of coverage site with at least 1× in reference genome. [e] The percentage of coverage site with at least 4× in reference genome.

3.2. Phylogenetic and Population Structure Analyses

PCA revealed a high degree of dispersion among the three groups (Figure 2). Together, the two first principal components explained 17.9% of the genetic variation in the 87 genotypes (PC1, 10.4%; PC2, 7.5%).

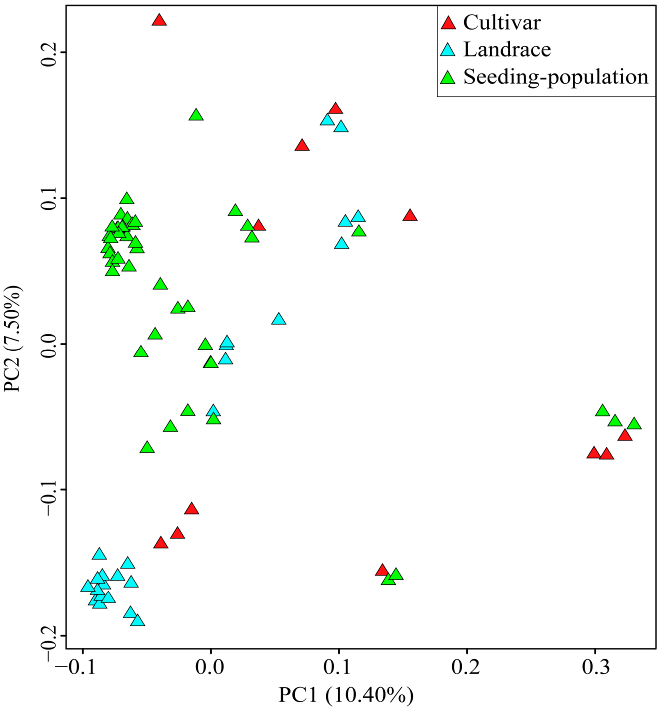

Figure 2. Principal component analysis (PCA) of 87 genotypes, based on 110,497 high-quality SNPs. Groupings that derive from structure analysis are coded by colors.

To investigate the evolutionary relationships among the 87 genotypes, a phylogenetic tree was constructed with MEGA 6 using high-quality SNPS (Figure 3). Phylogenetic analysis revealed three distinct groups. Group I was the largest, with 36 genotypes (35 seedling populations and 1 cultivar), and is hereafter referred to as 'Seedling-population'; Group II contained 22 genotypes (8 cultivars, 8 seedling populations, and 6 landraces), and is hereafter referred to as 'Cultivar'; group III contained 29 landraces (7 seedling populations, 3 cultivars, and 19 landraces), and is hereafter referred to as 'Landrace.' Most of the cultivars were mixed with seedling populations and landraces, such as LLZL, LD-Z, MJ-52, LCS, XSSR, and QBX-7, suggesting that these genotypes share the same origin.

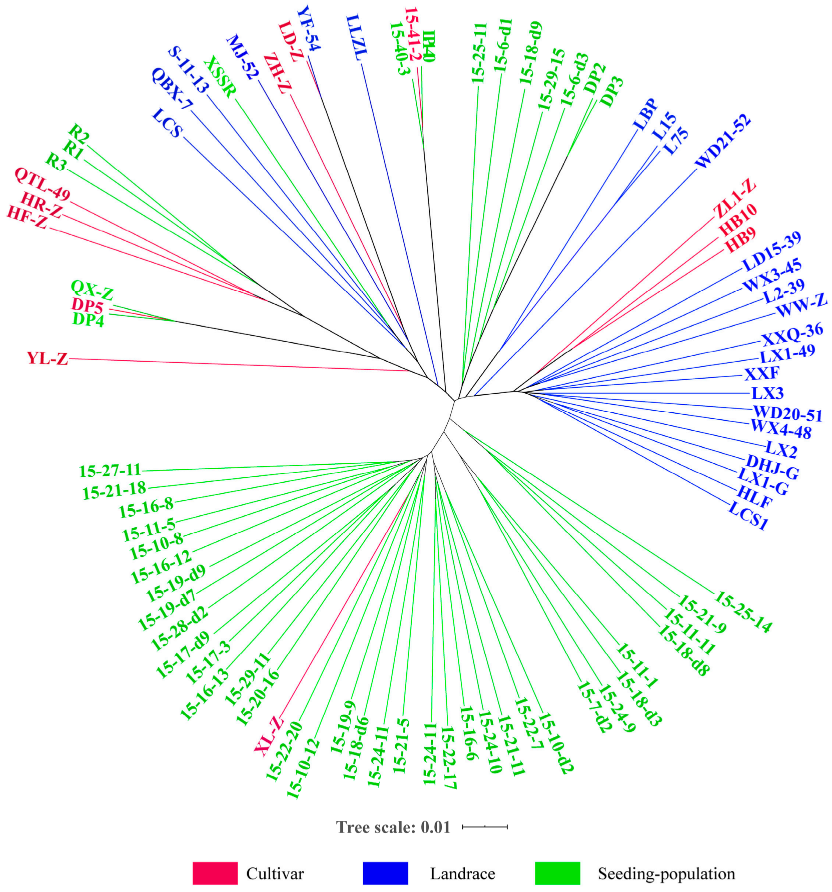

Figure 3. Phylogenetic analysis of 87 genotypes based on high-quality SNPs obtained by GBS. The neighbor-joining dendrogram was constructed with the maximum-likelihood method using MEGA 6. Green, blue, and red colors represent the seedling population, landrace, and cultivar groups, respectively.

To assess in greater detail the population structure of the genotype's diversity panel, the SNP genotype dataset was further analyzed and used for the construction of the phylogenetic tree through the model-based ADMIXTURE method carried out in STRUCTUR [37]. The optimal subgroup classification (K = 3) was inferred based on the lowest cross-validation error (Figures 4 and 5). In general, each population was represented by each species. Population analysis also divided the 87 genotypes into three groups, consistent with the results of PCA (Figure 2) and phylogenetic analysis (Figure 3).

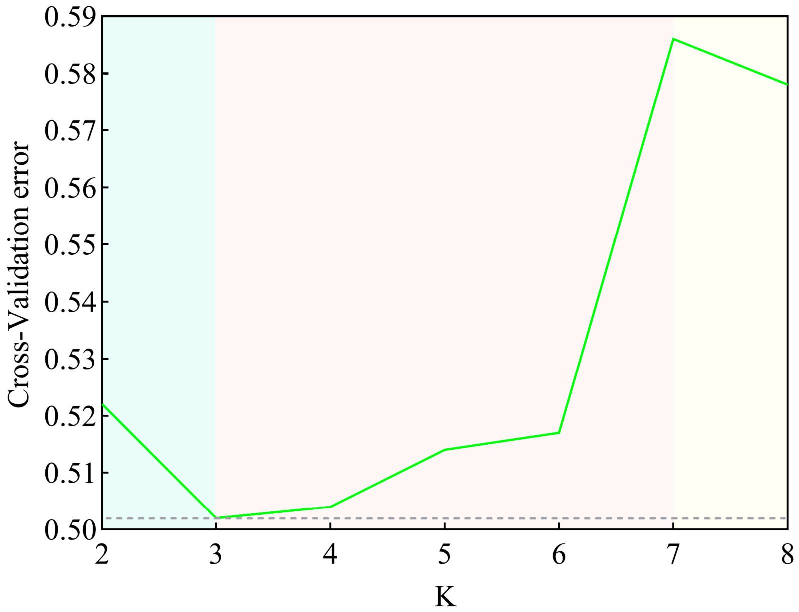

Figure 4. Cross-validation calculation for K ranging from 2 to 8.

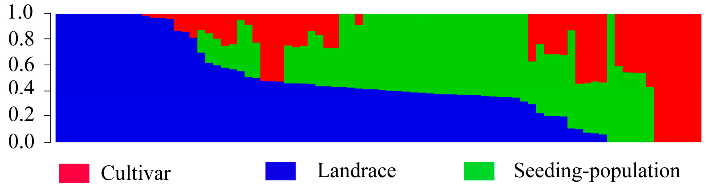

Figure 5. Genetic relationship among 87 genotypes surveyed using STRUCTURE version 2.3.4. K = 3 showed the lowest cross-validation error value. Different colors indicate different subpopulations. If a sample is composed of multiple colors, the proportion of each subpopulation can be seen, corresponding to the ordinate.

3.3. Construction of the Core Collection

A core collection is a subset of the germplasm representing the diversity of the original population and can be used for marker development and crop breeding. To establish the core collection, a subset of the 110,497 high-quality SNPs (6540 core SNPs) with 4× read depth, 80% minimum coverage across all 87 genotypes, MAF ≥ 20%, and polymorphism information content (PIC) ≥ 30% were selected (Supplemental Table S3). Of the 6540 core SNPs, 3407 (52.09%) were located in introns, 1289 (19.71%) in exons, 1407 (21.51%) in upstream or downstream regions, 432 (6.61%) in 5′- or 3′-untranslated regions (5′/3′UTRs), and 5 (0.08%) in splice site regions (Table 2). The core SNPs were distributed across all 16 chromosomes (Figure 6), with the highest number of chromosome 7; however, 13 core SNPs could not be assigned to a chromosome (Table 3).

Table 2. Genomic distribution of core SNPs.

Genomic Region	No. of Core SNPs	Relative Proportion of Core SNPs (%)
Introns	3407	52.09
Exons	1289	19.71
Splice sites	5	0.08
Upstream/downstream regions	1407	21.51
5′UTR/3′UTR	432	6.61
Total	6540	100

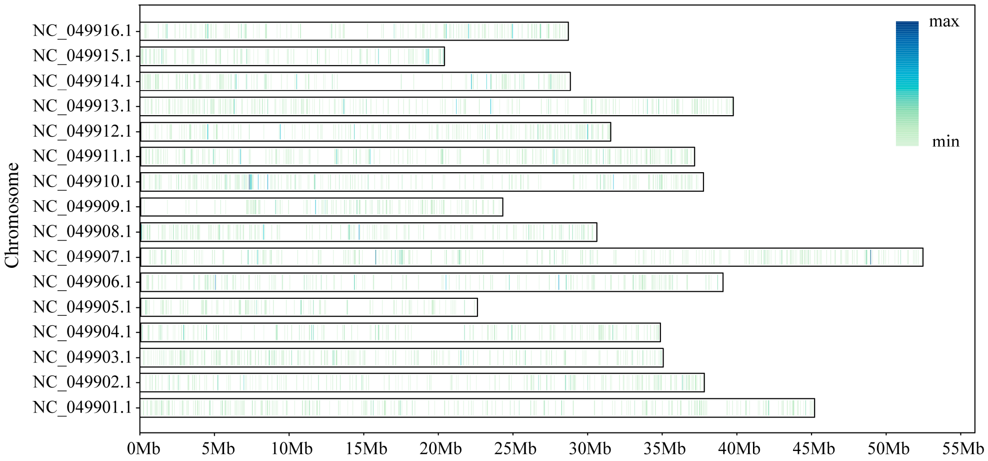

Figure 6. Distribution of core SNPs across 16 walnut chromosomes. The bottom scale represents the physical location of core SNPs on each chromosome. Color scale indicates the core SNP density (green, low density; blue, high density).

Table 3. Chromosomal distribution of core SNPs.

Chromosome No.	No. of Core SNPs	Relative Proportion of Core SNPs (%)
Chr 1	531	8.12
Chr 2	403	6.16
Chr 3	462	7.06
Chr 4	361	5.52
Chr 5	234	3.58
Chr 6	399	6.1
Chr 7	592	9.05
Chr 8	377	5.76
Chr 9	282	4.31
Chr 10	522	7.98
Chr 11	529	8.09
Chr 12	359	5.49
Chr 13	486	7.43
Chr 14	351	5.37
Chr 15	279	4.27
Chr 16	360	5.5
Unknown	13	0.2
Total	6540	100

Among the 87 genotypes, 9 were identified with Core Hunter 3.2.1, representing 10% of the entire allelic diversity in the collection (Supplemental Table S4), which is completely in accordance with a sampling scale of approximately 5–40%. These nine genotypes included 2 of the 36 seedling populations, 4 of the 22 cultivars, and 3 of the 29 landraces (Figure 2). Comparison of the observed heterozygosity (Ho), expected heterozygosity (He), and nucleotide diversity (π) revealed no significant difference between the primary set and core collection (Table 4).

Table 4. Statistics of genetic diversity in the primary set and core collection.

Pop	Ho	*p*	SE	He	*p*	SE	π	*p*	SE
All	0.4163	0.0138	0.0008	0.4208	0.0039	0.0004	0.4235	0.004	0.0004
Core	0.3309	0.0371	0.0014	0.3782	0.0121	0.0008	0.4104	0.0143	0.0009

Ho: average observed heterozygosity; He: average expected heterozygosity; π: nucleotide diversity; SE: standard error; *p*: *p*-value.

4. Discussion

Evaluation of genetic diversity is important for the genetic improvement of historically under-researched species, such as *J. regia*. GBS is an NGS-based technique that requires the analysis of only restriction enzyme sites in the genome sequence, thus allowing a quick, low-cost, and effective analysis of many samples [39,40]. However, the GBS method requires the establishment of a database based on enzyme digestion patterns, since the efficiency of enzyme digestion will significantly affect the quality of database construction and sequencing output, which can easily lead to data loss. In this study, more than 1.6 million high-quality SNPs were identified in 87 walnut genotypes. Among these SNPs, 6.84% (110,497 SNPs) passed the stringent filtering criteria (missing data < 20%, MAF > 5%); this percentage was much higher than the 0.71% reported previously [41]. One of the main reasons for the rapid decline in SNP number is the missing data, which often occurs in GBS-based research [42,43], although other factors may also be responsible, such as the incomplete Juglans reference genome sequence and short sequence alignment. Therefore, with a large sample size, deficiencies can be compensated by data filling, and the design is widely used in genome selection research [44,45].

The diversity panel of *J. regia* germplasm resources used in the current study was the result of collections made by multiple groups. The 87 walnut genotypes were classified into three major groups through phylogenetic analysis. In cluster analysis, more than half of the genotypes showed a certain correlation with their geographical origin, but a small number of mismatch phenomena were observed. For instance, three cultivars were placed into the landrace group, with a mismatch rate of 10.34. This was probably because these cultivars were selected from landraces or might have undergone gene flow. Thus, the three cultivars should belong to the landrace group. Similar results have been obtained in studies on other species [46]. In addition, the results of cluster analysis of cultivars provide interesting insights into their pedigrees. For example, the source country of some cultivars was different from their geographical origin; thus, our study provides new information regarding the plant material. Our results indicated that some cultivars were derived from crosses between genotypes from various countries of origin. Population structure analysis is a widely used method for inferring hidden population structure in plant species [47]. In the present study, population structure analysis and PCA confirmed the three major groups that did not correspond to geographical origin. A poor association between the molecular marker data and the geographic origin of genotypes has also been reported previously [30,48,49]. This may be caused by the extensive preservation and exchange of germplasm to broaden the diversity of local breeding materials, which increases the genetic similarity among materials in the local gene bank. Subsequently, after further analyzing pedigree, germplasm collection, growth habit, and various source, we found that pedigree was also the main factor for separation among the 87 walnut genotypes. This explains why cultivars were crossed with landraces. Our work shows that all seedlings are known hybrids. The pedigrees of 35 and 2 hybrids contained 'XL-Z' and 'LP1-40' cultivars, respectively. Additionally, six hybrids contained different landraces in their pedigrees. The surveyed landraces were collected in different years or from different climate zones and geographical regions, and grouped close together, reflecting a low level of genetic diversity. Therefore, the genetic diversity of landraces could be related to their adaptation to the local environment, and growers likely played a crucial role in maintaining the genetic diversity [50].

Core collections serve as vital resources in the germplasm resource bank and aim to conserve as much genetic diversity of the original collection as possible within a small number of genotypes [17]. The construction of a core collection can offer researchers and breeders options for protecting the genetic diversity of plant species over the long term. In recent years, with the progress in sequencing technology, molecular markers have been used for the development of core collections [21–23]. Obviously, molecular markers are more useful than morphological markers; molecular markers are based on DNA attributes, whereas morphological markers are frequently affected by epigenetic variation, which is no special consideration if DNA is used alone.

Many algorithms are available for the construction of a core collection, including PowerCore, Maximum Length Sub Tree (MLST), and Core Hunter [51]. Previous studies have shown that core collections of walnuts constructed using multiple algorithms are highly similar [52]. In contrast, core Hunter is a powerful and flexible algorithm for the selection of core genotypes, which exhibit high average genetic distance among germplasm resources and rich genetic diversity as a whole. The core subset of the samples diversifies large germplasm resources with minimal redundancy. Accordingly, a core collection of *J. regia* was developed in this study based on GBS data using the R package Core Hunter software [38]. The results of statistical analysis using the Student's t-test were consistent with the standards of the core collection. Assessment of initial and core collections by analysis of variance (ANOVA) revealed that all molecular genetic variation arose within the collections, and the two collections possessed similar genetic diversity. He reflects the richness and evenness of alleles within a population. In the present research, the value of He was greater than that of Ho, indicating high heterozygosity in the *J. regia* core collection. Overall, the constructed core collection represented the initial collection, thus validating its effectiveness.

5. Conclusions

This work demonstrates the power of the GBS approach for investigating the detailed population structure and genetic diversity of the local walnut germplasm resources and utilizing this information to establish a core collection. The walnut seedling populations, landraces, and cultivars collected from different geographical regions and in different years were divided into three distinct groups. These germplasms were observed to have differences in geographical origin and pedigree. The core collection identified from the germplasm collection will be useful for the rational and economically sustainable management of the walnut germplasm while helping to preserve its genetic diversity. This information can assist walnut breeders in designing more effective breeding programs for improving particular traits of interest, such as abiotic stress tolerance, nut and kernel quality, and disease resistance.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agronomy13071786/s1, Table S1: the information of 87 genotypes; Table S2: the output statistics of sequencing data; Table S3: the information of core SNP tag; Table S4: core collection sample.

Author Contributions: Conceptualization, J.R. and X.-Y.L.; methodology, J.R.; investigation, All authors; resources, All authors; data curation, J.R.; writing—original draft preparation, J.R.; writing—review and editing, J.R.; project administration, X.-Y.L., Y.-A.W. and X.-K.Z.; funding acquisition, J.R., X.-Y.L., Y.-A.W. and X.-K.Z.; sampling, J.R. and X.-Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (grant numbers 31860204 and 32260412); by the Gansu Provincial Natural Science Foundation Project 20JR5RA105 from the Science and Technology Department of Gansu Province; and by the discipline-team project of the Gansu Academy of Agricultural Sciences (grant number 2022-GAAS-07).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Abdallah, I.B.; Tlili, N.; Martinez-Force, E.; Rubio, A.G.P.; Perez-Camino, M.C.; Albouchi, A.; Boukhchina, S. Content of carotenoids, tocopherols, sterols, triterpenic and aliphatic alcohols, and volatile compounds in six walnuts (*Juglans regia* L.) varieties. *Food Chem.* **2015**, *173*, 972–978. [CrossRef] [PubMed]
2. Beer, R.; Kaiser, F.; Schmidt, K.; Ammann, B.; Carraro, G.; Grisa, E.; Tinner, W. Vegetation history of the walnut forests in Kyrgyzstan (central Asia): Natural or anthropogenic origin? *Quaternary Sci. Rev.* **2008**, *27*, 621–632. [CrossRef]

3. FAO. Food and Agriculture Organization of the United Nation Statistics for 2020. Available online: http://www.fao.org/faostat/en/#home (accessed on 10 September 2020).
4. Wang, G.Y.; Chen, Q.; Yang, Y.; Duan, Y.W.; Yang, Y.P. Exchanges of economic plants along the land silk road. *BMC Plant Biol.* **2022**, *22*, 619. [CrossRef] [PubMed]
5. Pollegioni, P.; Woeste, K.; Chiocchini, F.; Lungo, S.D.; Ciolfi, M.; Olimpieri, I.; Tortolano, V.; Clark, J.; Hemery, G.E.; Mapelli, S.; et al. Rethinking the history of common walnut (*Juglans regia* L.) in Europe: Its origins and human interactions. *PLoS ONE* **2017**, *12*, e0172541. [CrossRef] [PubMed]
6. Han, H.; Woeste, K.E.; Hu, Y.; Dang, M.; Zhang, T.; Gao, X.X.; Zhou, H.J.; Feng, X.J.; Zhao, G.F.; Zhao, P. Genetic diversity and population structure of common walnut (*Juglans regia*) in China based on EST-SSRs and the nuclear gene phenylalanine ammonia-lyase (PAL). *Tree Genet. Genomes* **2016**, *12*, 111. [CrossRef]
7. Ding, Y.M.; Cao, Y.; Zhang, W.P.; Chen, J.; Liu, J.; Li, P.; Renner, S.S.; Zhang, D.Y.; Bai, W.N. Population-genomic analyses reveal bottlenecks and asymmetric introgression from Persian into iron walnut during domestication. *Genomes Biol.* **2022**, *23*, 145. [CrossRef]
8. Shavvon, R.S.; Qi, H.L.; Mafakheri, M.; Fan, P.Z.; Wu, H.Y.; Vahdati, F.B.; Al-Shmgani, H.S.; Wang, Y.H.; Liu, J. Unravelling the genetic diversity and population structure of common walnut in the iranian plateau. *BMC Plant Biol.* **2023**, *23*, 201. [CrossRef]
9. Woodworth, R.H. Meiosis of micro-sporogenesis in the Juglandaceae. *Am. J. Bot.* **1930**, *17*, 863–869. [CrossRef]
10. Feng, X.; Zhou, H.J.; Zulfiqar, S.; Luo, X.; Hu, Y.H.; Feng, L.; Malvolti, M.E.; Woeste, K.; Zhao, P. The phytogeographic history of commom walnut in China. *Front. Plant Sci.* **2018**, *9*, 1399. [CrossRef]
11. Hu, Y.; Dang, M.; Feng, X.; Woeste, K.; Zhao, P. Genetic diversity and population structure in the narrow endemic Chinese walnut *Juglans hopeiensis* Hu: Implications for conservation. *Tree Genet. Genomes* **2017**, *13*, 91. [CrossRef]
12. Martínez-García, P.J.; Crepeau, M.W.; Puiu, D.; Gonzalez-Ibeas, D.; Whalen, J.; Stevens, K.A. The walnut (*Juglans regia* L.) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant Mol. Biol.* **2016**, *87*, 507–532. [CrossRef] [PubMed]
13. Aradhya, M.; Velasco, D.; Ibrahimov, Z.; Toktoraliev, B.; Preece, J.E. Genetic and ecological insights into glacial refugia of walnut (*Juglans regia* L.). *PLoS ONE* **2017**, *12*, e0185974. [CrossRef] [PubMed]
14. Germain, E.; Prunet, J.P.; Garcin, A. *Walnuts*; CTIFL: Paris, French, 1999.
15. Hassani, D.; Mozaffari, M.R.; Souraki, Y.D.; Soleimani, A.; Loni, A. Vegetative and reproductive traits of some Iranian local and foreign cultivars and genotypes of walnut (*Juglans regia* L.). *Seed Plant Improv. J.* **2013**, *29*, 839–855. [CrossRef]
16. Orhan, E.; Eyduran, S.P.; Poljuha, D.; Akin, M.; Ercisli, S. Genetic diversity detection of seed-propagated walnut (*Juglans regia* L.) germplasm from eastern Anatolia using SSR markers. *Folia Hortic.* **2020**, *32*, 37–46. [CrossRef]
17. Liu, Y.L.; Geng, Y.P.; Xie, X.D.; Zhang, P.F.; Hou, J.L.; Wang, W.Q. Core collection construction and evaluation of the genetic structure of glycyrrhiza in china using markers for genomic simple sequence repeats. *Genet. Resour. Crop Evol.* **2020**, *67*, 1839–1852. [CrossRef]
18. Magige, E.A.; Fan, P.Z.; Wambulwa, M.C.; Milne, R.; Wu, Z.Y.; Luo, Y.H.; Khan, R.; Wu, H.Y.; Qi, H.-L.; Zhu, G.-F.; et al. Genetic Diversity and Structure of Persian Walnut (*Juglans regia* L.) in Pakistan: Implications for Conservation. *Plants* **2022**, *11*, 1652. [CrossRef]
19. Frankel, O.H.; Brown, A.H.D. Plant genetic resources today: A critical appraisal. In *Crop Genetic Resources: Conservation and Evaluation*; Holden, J.H.W., Williams, J.H., Eds.; George Allan and Unwin: London, UK, 1984; pp. 249–257.
20. Bhattacharjee, R.; Khairwal, I.S.; Bramel, P.J.; Reddy, K.N. Establishment of a pearl millet [*Pennisetum glaucum* (L.) R. Br.] core collection based on geographical distribution and quantitative traits. *Euphytica* **2007**, *155*, 35–45. [CrossRef]
21. Kumar, S.; Ambreen, H.; Variath, M.T.; Rao, A.R.; Agarwal, M.; Kumar, A.; Goel, S.; Jagannath, A. Utilization of molecular, phenotypic, and geographical diversity to develop compact composite core collection in the oilseed crop, safflower (*Carthamus tinctorius* L.) through maximization strategy. *Front. Plant Sci.* **2016**, *19*, 1554. [CrossRef]
22. Mahmoodi, R.; Dadpour, M.R.; Hassani, D.; Zeinalabedini, M.; Vendramin, E.; Micali, S.; Nahandi, F.Z. Development of a core collection in Iranian walnut (*Juglans regia* L.) germplasm using the phenotypic diversity. *Sci. Hortic.* **2019**, *249*, 439–448. [CrossRef]
23. Poland, J.A.; Rife, T.W. Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* **2012**, *5*, 92–102. [CrossRef]
24. Belaj, A.; Domínguez-García, M.C.; Atienza, S.G.; Martín Urdíroz, N.; Rosa, R.D.; Šatović, Z.; Martín, A.; Kilian, A.; Trujillo, I.; Valpuesta, V.; et al. Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet. Genomes* **2012**, *8*, 365–378. [CrossRef]
25. Dhanaraj, A.L.; Rao, E.B.; Swamy, K.R.; Bhat, M.G.; Prasad, D.T.; Sondur, S.N. Using RAPDs to assess the diversity in Indian cashew (*Anacardium occidentale* L.) germplasm. *J. Hortic. Sci. Biotechnol.* **2002**, *77*, 41–47. [CrossRef]
26. García-Lor, A.; Luro, F.; Ollitrault, P.; Navarro, L. Comparative analysis of core collection sampling methods for mandarin germplasm based on molecular and phenotypic data. *J. Hortic. Sci. Biotechnol.* **2017**, *171*, 327–339. [CrossRef]
27. Ebrahimi, A.; Zarei, A.; Zamani Fardadonbeh, M.; Lawson, S. Evaluation of genetic variability among "Early Mature" *Juglans regia* using microsatellite markers and morphological traits. *PeerJ* **2017**, *5*, e3834. [CrossRef] [PubMed]
28. Zhu, T.; Wang, L.; You, F.M.; Rodriguez, J.C.; Deal, K.; Chen, L.; Li, J.; Chakraborty, S.; Balan, B.; Jiang, C.; et al. Sequencing a *Juglans regia* × *J. microcarpa* hybrid yields high-quality genome assemblies of parental species. *Hortic. Res.* **2019**, *6*, 1–16. [CrossRef]

29. Xu, Z.; Hu, T.; Zhang, F. Genetic diversity of walnut revealed by AFLP and RAPD markers. *J. Agric. Sci.* **2012**, *4*, 271–276. [CrossRef]
30. Mahmoodi, R.; Dadpour, M.R.; Hassani, D.; Zeinalabedini, M.; Vendramin, E.; Leslie, C.A. Composite core set construction and diversity analysis of Iranian walnut germplasm using molecular markers and phenotypic traits. *PLoS ONE* **2021**, *16*, e0248623. [CrossRef]
31. Li, H.; Durbin, R. Fast and accurate short read alignmeng with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
32. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
33. Tamura, A.; Urakami, H.; Tsuruhara, T. Purification of rickettsia tsutsugamushi by percoll density gradient centrifugation. *Microbiol. Immunol.* **1982**, *26*, 321–328. [CrossRef]
34. Perrière, G.; Gouy, M. WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie* **1996**, *78*, 364–369. [CrossRef]
35. R Core Team. A language and environment for statistical computing. *Computing* **2014**, *14*, 12–21.
36. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inferences of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945–959. [CrossRef] [PubMed]
37. Alexander, D.H.; Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform.* **2011**, *12*, 246. [CrossRef] [PubMed]
38. Beukelaer, H.D.; Davenport, G.F.; Fack, V. Core Hunter 3: Fast and Flexible Core Subset Selection. *BMC Bioinform.* **2018**, *19*, 203. [CrossRef]
39. Davey, J.W.; Hohenlohe, P.A.; Etter, P.D.; Boone, J.Q.; Catchen, J.M.; Blaxter, M.L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **2011**, *12*, 499–510. [CrossRef] [PubMed]
40. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **2011**, *6*, e19379. [CrossRef]
41. Zhao, P.; Zhou, H.J.; Potter, D.; Hu, Y.H.; Woeste, K. Population genetics, phylogenomics and hybrid speciation of Juglans in China determined from whole chloroplast genomes, transcriptomes, and genotyping-by-sequencing (GBS). *Mol. Phylogenet. Evol.* **2018**, *126*, 250–265. [CrossRef] [PubMed]
42. Elbasyoni, I.; Lorenz, A.J.; Guttieri, M.; Frels, K.; Baenziger, P.S.; Poland, J.; Akhunov, E. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci.* **2018**, *270*, 123–130. [CrossRef] [PubMed]
43. Wu, Y.; Vicente, F.S.; Huang, K.; Dhliwayo, T.; Costich, D.E.; Semagn, K.; Sudha, N.; Olsen, M.; Prasanna, B.M.; Zhang, X. Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theor. Appl. Genet.* **2016**, *129*, 753–765. [CrossRef]
44. Davies, R.W.; Flint, J.; Myers, S.; Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **2016**, *48*, 965–969. [CrossRef] [PubMed]
45. Chan, A.W.; Hamblin, M.T.; Jannink, J.L. Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data. *PLoS ONE* **2016**, *11*, e0160733. [CrossRef] [PubMed]
46. Jiang, X.B.; Fang, Z.; Lai, J.S.; Wu, Q.; Wu, J.; Gong, B.C.; Wang, Y.P. Genetic Diversity and Population Structure of Chinese Chestnut (*Castanea mollissima* Blume) Cultivars Revealed by GBS Resequencing. *Plants* **2022**, *24*, 3524. [CrossRef] [PubMed]
47. Xiao, Y.J.; Cai, D.F.; Yang, W.; Ye, W.; Younas, M.; Wu, J.S.; Liu, K. Genetic structure and linkage disequilibrium pattern of a rapeseed (*Brassica napus* L.*)* association mapping panel revealed by microsatellites. *Theor. Appl. Genet.* **2012**, *125*, 437–447. [CrossRef] [PubMed]
48. Sreekanth, P.M.; Balasundaran, M.; Nazeem, P.A.; Suma, T.B. Genetic diversity of nine natural *tectona grandis* L.f. populations of the western ghats in southern India. *Conserv. Genet.* **2012**, *13*, 1409–1419. [CrossRef]
49. Dangl, G.S.; Woeste, K.; Aradhya, M.K.; Koehmstedt, A.; Simon, C.; Potter, D. Characterization of 14 microsatellite markers for genetic analysis and cultivar identification of walnut. *J. Am. Soc. Hortic. Sci.* **2005**, *130*, 348–354. [CrossRef]
50. Zeven, A.C. Traditional maintenance breeding of landraces: Practical and theoretical considerations on maintenance of variation of landraces by farmers and gardeners. *Euphytica* **2002**, *123*, 147–158. [CrossRef]
51. Odong, T.L.; Jansen, J.; Van Eeuwijk, F.; Van Hintum, T.J. Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theor. Appl. Genet.* **2013**, *126*, 289–305. [CrossRef]
52. Bernard, A.; Barreneche, T.; Donkpegan, A.; Lheureux, F.; Dirlewanger, E. Comparison of structure analyses and core collections for the management of walnut genetic resources. *Tree Genet. Genomes* **2020**, *16*, 76. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.