forests

MDPI

Article

Comparative Analysis of SNP Discovery and Genotyping in *Fagus sylvatica* L. and *Quercus robur* L. Using RADseq, GBS, and ddRAD Methods

Bartosz Ulaszewski *, Joanna Meger and Jaroslaw Burczyk *

Department of Genetics, Faculty of Biological Sciences, Kazimierz Wielki University, Chodkiewicza 30, 85-064 Bydgoszcz, Poland; warmbier@ukw.edu.pl

* Correspondence: ulaszewski@ukw.edu.pl (B.U.); burczyk@ukw.edu.pl (J.B.)

Abstract: Next-generation sequencing of reduced representation genomic libraries (RRL) is capable of providing large numbers of genetic markers for population genetic studies at relatively low costs. However, one major concern of these types of markers is the precision of genotyping, which is related to the common problem of missing data, which appears to be particularly important in association and genomic selection studies. We evaluated three RRL approaches (GBS, RADseq, ddRAD) and different SNP identification methods (de novo or based on a reference genome) to find the best solutions for future population genomics studies in two economically and ecologically important broadleaved tree species, namely *F. sylvatica* and *Q. robur*. We found that the use of ddRAD method coupled with SNP calling based on reference genomes provided the largest numbers of markers (28 k and 36 k for beech and oak, respectively), given standard filtering criteria. Using technical replicates of samples, we demonstrated that more than 80% of SNP loci should be considered as reliable markers in GBS and ddRAD, but not in RADseq data. According to the reference genomes' annotations, more than 30% of the identified ddRAD loci appeared to be related to genes. Our findings provide a solid support for using ddRAD-based SNPs for future population genomics studies in beech and oak.

Keywords: restriction-site associated DNA sequencing; SNP genotyping; technical replicates; population genomics; European beech; pedunculate oak

Citation: Ulaszewski, B.; Meger, J.; Burczyk, J. Comparative Analysis of SNP Discovery and Genotyping in *Fagus sylvatica* L. and *Quercus robur* L. Using RADseq, GBS, and ddRAD Methods. *Forests* **2021**, *12*, 222. https://doi.org/10.3390/f12020222

Academic Editor: Tadeusz Malewski
Received: 26 January 2021
Accepted: 11 February 2021
Published: 15 February 2021

1. Introduction

In the past decade, the development of the next-generation sequencing (NGS) methods combined with various types of newly developed genomic library preparation protocols provided the tools for relatively inexpensive discovery and genotyping of large numbers of loci useful in population genomics studies [1–4]. The ultimate way of obtaining genomic data from multiple samples is to apply whole-genome sequencing (WGS). This approach maximizes the quantity of information gathered, and opens up the possibility of a wide variety of analyses; however, it is currently prohibitively expensive and computationally challenging [5], especially in non-model species with large genomes. Analyses of genomic variations relevant in most population genomic studies can be conducted with reduced representation genomic libraries (RRL) [6,7] where only a few percent of the genome is sequenced. The most popular techniques use restriction enzymes to prepare the DNA (restriction-site-associated DNA: RAD) for sequencing. In recent years many methods based on the RAD approach were developed, differing in the number of enzymes used or in additional steps of library preparation [8].

In this study, we have focused on three RRL methods: RADseq [9], GBS (genotyping-by-sequencing) [10], and ddRAD (double-digest RAD sequencing) [11–13]. In RADseq, DNA is digested with a single, frequently cutting restriction enzyme. To such prepared DNA, barcodes and common adaptors with the first of two primer (P1) are ligated, then samples are pooled, randomly sheared, size-selected within a 300–700 bp window and,

finally, in the final step P2 adaptors are ligated. Only the DNA containing both adaptors is PCR-amplified and sequenced in the single-end mode [9]. GBS is a simplified protocol also relying on a single restriction enzyme. Barcoded adaptors and common adaptors are randomly ligated to digested DNA, and fragments from multiple samples are pooled. Short DNA reads with both adaptor types are amplified and sequenced in single-end mode [10]. The ddRAD protocol uses two restriction enzymes—a rare and a frequent cutter. The first is used to generate fragments, after which barcoded primers are annealed. The second enzyme is used as the replacement for random shearing to improve the size selection step. Finally, P2 primers are ligated, and the fragments are amplified and sequenced afterward in double-end mode [11].

Precision and repeatability of genomic data are essential in most types of population genomics analyses, particularly in genomic selection surveys [14]. However, the ability to accurately genotype SNP loci from restriction-enzyme-based sequencing is a major concern [8]. Problems occurring in the genotyping step can be categorized into two groups: missing data and genotyping errors. These problems may originate from incorrect library preparation and un-optimized bioinformatics processing [13]. When ignored, these biases might affect the inferences of downstream analyses, to an unpredictable degree [15,16]. Optimizing library preparation and sequencing [13] mitigates sequencing errors. Unfortunately, some types of problems are unavoidable with restriction-enzyme-based methods, especially when the DNA quality is low [8,15].

On the other hand, sequencing artifacts can be minimized to some degree at the data preparation and core bioinformatics stage [17–19]. Current bioinformatics tools for genotyping from RAD data follow a series of three crucial steps: (1) raw data processing (preparation); (2) reading of the alignment against a reference genome or de novo assembly of the sequence tags; and (3) variant calling and filtering [13,20]. These steps aim to provide reliable and precise data; however, due to the probabilistic nature of the bioinformatic algorithms, they can alleviate errors while also generating some new mistakes that can have a profound effect on the final results of downstream analyses [21]. However, overly conservative filtering of genomic SNP data may cause data loss leading to misestimation of genetic effects [12].

One of the main advantages of using RAD methods for genotyping of non-model species is that loci can be identified de novo, without the availability of a reference genome [8]. However, the presence of reference genomes may be beneficial as RAD loci predicted in combination with a reference genome will be useful in filtering SNPs from paralogous or repetitive sequences, identifying indel variation, and avoidance of calling wrong loci due to biological contaminations [8,22]. With well assembled and annotated reference genomes, identified RAD loci can be readily positioned along the genome and may become directly useful in association studies. However, only a few studies have compared the efficiency and precision of SNP discovery and genotyping based on de novo versus reference-aligned approaches to date [23–25].

Regardless of the nature of any genotyping problems, one way to monitor the levels of inconsistencies in SNP discovery and genotyping is to use technical replicates [13,19] or include parent-offspring dyads, if permitted by experimental designs [23]. Mastretta-Yanes et al. [19] defined several types of possible errors that could be investigated based on the use of technical replicates while varying different parameters in Stacks software [26]. Using ddRAD libraries, they analyzed a non-model plant species without a reference genome. The study demonstrated that with technical replicates, it is possible to optimize and tune a de novo genotyping pipeline and to identify and mitigate sources of errors [19].

Due to their foundation roles in terrestrial ecosystems and their broad economic importance, forest tree species have been thoroughly investigated in the areas of population genetics and evolutionary biology. The accumulated genomic resources, including genome assemblies of major tree taxa (*Populus trichocarpa* [27], *Eucalyptus grandis* [28], *Pinus taeda* [29], *Olea europea* [30], and *Quercus lobata* [31]), accelerated the progress of population genomics in forest trees. The number of studies involving genomic data is continuously increasing, including works based on RAD approaches [22]. Forest trees, due to their characteristic

life traits [32] are highly heterozygous as compared to other species [33]. The high level of genome-wide heterozygosity may complicate genome assemblies [31] and confound SNP discovery and genotyping in RAD-based experiments. However, studies aimed at optimizing RAD approaches in forest trees are limited [22].

In this study, we investigated the efficiency of SNP discovery and genotyping in two of the most important broadleaved tree species in Europe, namely common beech (*Fagus sylvatica* L.; abbr. FS) and pedunculate oak (*Quercus robur* L.; abbr. QR), both belonging to the Fagaceae family. Using the same set of individuals within the species, we evaluated the three RAD approaches mentioned earlier: RADseq, GBS, and ddRAD. Taking advantage of the reference genomes of beech and oak, we contrasted de novo and reference-aligned marker discovery approaches. Among the samples of each species, we included technical replicates (four individuals sampled twice), which enabled us to monitor the replicates genotyping consistency while optimizing specific parameters of the applied bioinformatics pipelines. Our ultimate goal was to fine-tune protocols and the approaches for gathering as many loci as possible with the fewest possible artifacts and the lowest proportion of missing loci. We believe that our findings will be useful for selecting the most appropriate RAD approaches for population studies in beech and oak, and will provide best-practice guidelines for processing RAD data in general.

2. Materials and Methods

2.1. Species Background

Common beech (*Fagus sylvatica* L.) and pedunculate oak (*Quercus robur* L.) are broad-leaved, monoecious, wind-pollinated, and highly outcrossed tree species [34,35], both belonging to the Fagaceae family. In Europe, they play essential roles in forest ecosystems' ecology and are essential in the forestry-based economy in several countries [36–38]. Genetics is one of the most thoroughly researched aspects of beech and oak [39,40]. However, while it is continuously growing, the genomic-based knowledge of these two species is still limited but continuously growing [41–45]. Reference genomes of the two species were published relatively recently. The common beech haploid genome (version: 1.3) [46] is the smaller of the two, totaling 542.3 Mbp with a GC-pair content of 35.69% and 62,085 predicted genes. The pedunculate oak haploid genome (version: PM1N) [47] is distinctly larger: 814.36 Mbp, consisting of 35.65% GC-pairs, and 25,808 well-defined genes.

2.2. Sample Collection

Beech individuals selected for this study were sampled from the provenance trial located in the Siemianice Experimental Forest District (51°12′52.4″ N, 17°59′50.1″ E) in south-central Poland [48]. The trial consists of 71 provenances originating mostly from Central Europe; however, for this study, we sampled 91 individuals representing 47 provenances. Among them, four individuals from four different provenances were sampled twice from the same individual and were considered technical replicates (duplicated samples) and marked thereafter as FA, FB, FC, and FD. Oak individuals originated from the provenance/family of a common-garden trial located in the Mogilica Forest District (53°12′36.7″ N, 15°13′39.7″ E) in north-western Poland [49]. The trial consists of the progeny of mother trees originating from eight Polish provenances (50 mother trees/provenance). As with beech, we sampled 91 individuals representing all of the provenances and 64 unique mother-trees families, and four individuals were selected as technical replicates, each one being the progeny of a different provenance/mother-tree, marked as QA, QB, QC, and QD. Such a wide sampling of individuals from several populations was intended to account for the possibility of large genetic diversity within species.

Replicates were processed separately to assess the repeatability of SNP identification and calling procedures. Altogether, within each species and method, 95 samples were subjected to DNA isolation and sequencing (570 samples in total). Samples for GBS and RADseq libraries were collected in 2014 from the same individuals. However, to generate ddRAD libraries, the trees were sampled in 2015, and some individuals (not the replicates)

were replaced due to health-related dropout cases. Details regarding the sampled material is presented in Tables S1.1 (for beech) and S1.2 (for oak).

2.3. DNA Isolation, Library Construction, and Sequencing

The sequencing of reduced-representation libraries demands high molecular weight DNA >50 kb, with a concentration >30 ng/µL and total mass of 1.5–3 µg [50]. To meet these requirements, we collected flushing leaf buds. Sampled material was dried to ≈10% humidity with a phytotron (BINDER WTC KB 240) for 24 h at 30 °C; afterwards, 30 mg of tissue was stacked in sterile 2 mL Eppendorf tubes and ground at 30 Hz in a laboratory mill (Mixer Mill MM 400, Retsch, Haan, Germany). DNA was isolated using GeneMATRIX Plant & Fungi DNA Purification Kit (EURx, Gdańsk, Poland), with slight modifications to the manufacturer's protocol (for details, see Supplementary Materials).

Isolated DNA was shipped to external service providers for enzyme optimization, library construction, and sequencing: GBS—Cornell University (Ithaca, NY, USA); RADseq—Floragenex Inc. (Portland, OR, USA); and ddRAD—IGA Technology Services (Udine, Italy). Details of library construction and sequencing are presented in Table 1.

Table 1. Summary of enzymes and sequencing modes used for each library type.

Library Type	Enzyme and Species	Sequencer	Read Length	Sequencing Mode	Barcode Length (bp)
GBS	*EcoT22I*—both	HiSeq 2000	100	Single-end	5–8
RADseq	*PstI*—both	HiSeq 2000	100	Single-end	10
ddRAD	*SphI/Sau3AI—F. sylvatica* *PstI/Sau3AI—Q. robur*	HiSeq 2500	125	Paired-end	15

2.4. Data Processing and Analysis

Sequence datasets were checked with FastQC software (v0.11.8; [51]) to assess the average Phred quality score of each nucleotide position, and the presence of artifacts such as sequencing adaptors. Samples were demultiplexed using the *process_radtags* tool: a part of the Stacks pipeline (v2.3) [26]. *Process_radtags* was configured to discard reads with average Phred-score <20, remove the barcode and trim 3′-ends of the reads if necessary. Among reads from the GBS library, the presence of contaminants (e.g., Illumina universal adaptors) and a quality drop at the 3′-end meant that these samples were trimmed to 64 bp. RADseq and ddRAD data were free from these artifacts, providing 90 bp and 110 bp reads, respectively. Quantity and quality of reads of each individual were again checked with FastQC; the results were summarized with MultiQC (v1.7) [52]. Finally, the reads were processed for additional purification with Trimmomatic [53] using default settings.

Initial analyses were conducted based on technical replicates. The intention behind the use of technical replicates was to investigate to what degree the genotypes reported from two replicates of the same individual are identical and how the parameters used in de novo or reference-based SNP calling affect that similarity. For the de novo approach, we used Stacks (v2.3) [26] to construct loci and extract SNPs. In this study, we tested a set of three parameters, similarly to the study of Mastera-Yanes et al. [19]. The core parameters in the *denovo_map* pipeline were varied across the course of the experiment: *m*—a minimum number of reads necessary to create a stack (range 2 ÷ 15, default 3); *M*—maximum number of mismatches between stacks while searching for the allele in an individual (range 2 ÷ 10, default 2); and *n*—the maximum number of mismatches between loci searched in a joint pool (range 1 ÷ 5, default 1). Only one parameter at a time was changed, while others were set to the default values. The results were outputted to 'vcf' file format using the built-in 'populations' pipeline, providing 28 datasets for each species and the library for further analyses (168 datasets in total).

In the reference-based approach, we used bwa software (v 0.7.17) [54] with default settings to map the reads against available reference genomes and tested the effects of varying different parameters on the outcome of the SNP-calling procedure. This step was carried out using Heap software (v0.8) [55]. The program uses an approach similar to

GATK [56], but it is less stringent and has a rather simpler setting. The following parameters were tested: *depth*—a minimum depth of filtered reads that support each of the reported alleles at the individual level (range 3 ÷ 10, default 3), and *mapq*—the posterior probability that mapping position is wrong, expressed as a Phred score (20, 30, 40; default 20). Similarly to the method outlined above for de novo analyses, only one parameter was changed at a time while other parameters were fixed to the default values. The results were outputted to a *vcf* file format, delivering 34 datasets for each species and the library for further analyses (234 datasets in total).

The results obtained for technical replicates were filtered with vcftools (v0.1.16) [57] to exclude indels and include only biallelic SNPs present in at least 6 out of 8 samples (75%). The key indicators and error types (Table 2) were calculated for each replicated pair in each dataset, and averaged with a custom-designed script written in bash. Note that due to filtering out of anything other than biallelic loci, the allele errors of type A | C-A | T (as indicated in [19]) could not be assessed in our study.

Table 2. Key indicators used to assess method efficiency.

Indicator	Description	Examples *
Good loci (GL)	Genotypes in both replicates are the same.	R_1 A | A—R_2 A | A R_1 A | C—R_2 A | C
Missing allele (MA)	A variant of one genotype partially fits the other	R_1 A | A—R_2 A | C
Locus error (LE)	Both genotypes differ with no common alleles	R_1 A | A—R_2 C | C
Missing loci (ML)	One genotype available, second is absent.	R_1 A | A—R_2 0 | 0
Missing data (MD)	Both genotypes of a replicate are absent.	R_1 0 | 0—R_2 0 | 0

* R_1/R_2—first and second technical replicates for a pair; A,C—example nucleotides; 0 | 0—missing genotype.

Finally, we analyzed the full dataset to assess the efficiency of different RRL libraries used in this study. To reduce redundancy in the full datasets, we have included only one replicate from a pair: the one with the higher read number. Samples with less than 50% of the average number of reads were discarded from the analysis [19]. SNPs from de novo (Stacks) and reference-based (Heap) protocols were generated using default settings. The outputted results were filtered with vcftools [57] to exclude indels, non-biallelic SNPs, and markers present in less than 80% of the samples. To reduce linkage disequilibrium (LD) bcftools was used with $r^2 > 0.5$ in a 1000 bp window. Filtered SNPs were analyzed to determine differences in the generated data from the GBS, RADseq, ddRAD genomic libraries and how genotyping strategies affect each of them.

Basic statistics of different datasets were generated with bcftools from the samtools package (v1.9) [58] with option: *stats*. A summary of the statistics was conducted in the R environment [59]. Annotation of the reference-based SNPs was carried out using SnpEff [60] using default settings, with gff3 files provided with the reference sequences of beech and oak [46,47]. The annotation was performed on the sets containing biallelic SNPs with markers present in at least 80% of the samples, and minor allele frequency (MAF) > 0.05.

3. Results

3.1. Technical Replicates

The highest average number of reads after cleaning was obtained for GBS libraries for both species: FS: 1.93×10^6 reads/sample; QR: 3.34×10^6 reads/sample. In the other two libraries the quantity of data was considerably lower: RADseq—FS 1.15×10^6 reads/sample, QR: 1.15×10^6 reads/sample; ddRAD—FS: 1.07×10^6 reads/sample; QR: 1.08×10^6 reads/sample (for details, see Table S2). We noticed that the replicated pairs with higher read numbers generated more SNPs; however, larger differences in the number of reads between individuals within the pair negatively affected the quantity of data obtained per pair.

The analyses conducted using default settings showed that different types of libraries provided comparable results for oak and beech (Table 3). The ddRAD and RADseq datasets, in contrast to GBS, generated a proportionally higher number of raw SNPs in the mapping

approach than in the de novo approach. We also found that after filtering (only biallelic SNPs; present in at least 6 out of 8 samples; indels removed), the set of SNPs generated de novo had a distinctly higher percentage of the retained markers than the set obtained using the mapping approach (average 53.4% vs. 43.2%; Table 3). However, the actual number of SNP loci after filtering was the highest for ddRAD, and from the mapping approach in particular (Table 3).

The proportion of good loci (GL) was significantly higher in GBS and ddRAD datasets as compared to RADseq (Table 3). This may have been caused by fragmention of DNA, as the RADseq protocol heavily relies on initial DNA quality and may produce uneven coverage among the samples [15,61]. This assumption is supported by the distinctly altered proportions of missing loci (ML) and allele error (LE) counts in the RADseq data as compared to the other two libraries. The proportion of missing alleles (MA) in the GBS and ddRAD sets increased when using the mapping method. This is probably a result of the SNP calling procedure, where the allele must be present in at least three reads to call a variant. In general, all libraries had similar levels of missing data (MD).

In conclusion, given the yield of filtered SNP loci and the proportion of reliable loci (GL) for both species, the most efficient method to gather the largest quantity of data is by applying the ddRAD technique and implementation of a mapping approach for SNP discovery and calling. Finally, in respect of 'good' loci, we were able to find 52,280 in beech and 67,977 in oak, given the filtering criteria applied.

3.2. Influence of Parameters Used in Genotyping Procedures

Finding the best-fitting parameters is a crucial aspect of bioinformatics [62]. By increasing the threshold for the minimum number of reads required to create a locus (*m*) in the de novo procedure using Stacks (v2.3) [26], the number of SNPs in all analyzed datasets was decreased. The parameter itself may be treated as an additional filtering tool for obtaining loci with higher coverage, and for increasing reliability of the results. Comparing the number of markers generated with a variable *m* value (for details see: Tables S3.1 and S3.2) indicated a decrease of SNPs reported in raw and filtered datasets, as expected. The share of good loci (GL) in all analyzed cases decreased with increasing *m*, which was correlated with the increase of missing loci (ML). In general, the *m* parameter had a minor effect on MA, LE, and MD; however, the filtering nature of the *m* value mitigated locus error (LE) to some degree in RADseq sets (for detailed results, see: Tables S3.1 and S3.2). Overall, we found the default value (3) of the *m* parameter to be the best in optimizing the quantity and quality of discovered SNPs for both species.

We found that the two Stacks parameters that allow for mismatches, namely *M* (number of mismatches between stacks within the individual) and *n* (number of mismatches in the joint pool of individuals), elevated the numbers of raw markers in all studied cases. Only in the QR-GBS dataset did the number of filtered SNPs slightly decreased with increasing *M* (Table S3.2). This occurs when a locus is underrepresented in a joint catalog being filtered for a minimum number of individuals, which causes a marker dropout. All datasets were insensitive in terms of the influence of the *M* parameter on the key indicators defined in Table 2. For example, we observed only a slight decrease (2.3%–4.9%) in the GL when increasing the number of mismatches (for detailed results, see: Tables S3.1 and S3.2). In general, our results indicated that the elevated number of mismatches within the individual has only a slight influence on the key indicators, and the most important factor responsible for marker quality is the initial quality of the data itself. However, when choosing the optimal value of *M*, several other factors must be considered, such as the length of sequence reads, the biology of species (including heterozygosity levels), and the degree of relatedness/differentiation among individuals or populations. Considering the above, and the fact that our sampling material in each species represented several populations (see Methods) we decided to choose the optimal value of the *M* value to be 2 (default), both for beech and oak, although slightly higher values should provide more markers with only a slight loss of GL.

Table 3. Total number of raw (unfiltered) and filtered SNPs, and key indicators for replicated samples and each library type and genotyping method; data generated with default settings (GL—good loci, MA—missing allele, LE—locus error, ML—missing loci, MD—missing data).

		Fagus sylvatica						***Quercus robur***					
		GBS		**RADseq**		**ddRAD**		**GBS**		**RADseq**		**ddRAD**	
	Average number of reads/sample	1,927,317		1,159,360		1,071,369		3,342,097		1,513,053		1,086,261	
	Genotyping method	de novo	map	de novo	map	de novo	map	de novo	map	de novo	map	de novo	map
	Number of raw SNPs	32,180	21,860	37,063	43,538	81,338	145,548	67,123	51,747	45,026	88,151	76,376	162,674
	Number of SNPs after filtering *	20,813	12,666	16,454	13,764	42,732	65,082	42,859	29,482	10,888	13,415	51,814	78,786
	% of SNPs after filtering *	64.68	57.94	44.39	31.61	52.54	44.72	63.85	56.97	24.18%	15.22%	67.84	48.43
Key indicators of the filtered SNPs	GL	87.01%	84.80%	46.23%	48.43%	80.90%	80.33%	87.50%	86.70%	42.60%	46.00%	87.25%	86.28%
	MA	2.04%	5.32%	17.21%	8.83%	2.80%	5.35%	2.0%	4.23%	15.1%	7.3%	1.86%	4.34%
	LE	0.12%	0.04%	5.45%	9.17%	0.18%	0.04%	0.2%	0.02%	6.8%	9.7%	0.10%	0.02%
	ML	6.94%	6.91%	28.89%	31.14%	12.07%	10.50%	6.1%	6.00%	32.9%	34.1%	7.03%	5.62%
	MD	3.89%	2.94%	2.22%	2.43%	4.04%	3.79%	4.2%	3.03%	2.6%	2.8%	3.75%	3.73%

* Only biallelic SNPs, present in at least 6 out of 8 samples, indels removed.

The *n* parameter is responsible for allowing mismatches in a joint pool, and as expected, increasing *n* resulted in increased numbers of SNPs in both raw and filtered data. The GL parameter decreased slightly in GBS and ddRAD sets (<5%) and was rather stable in RADseq. The effect of *n* on other key parameters was considerably lower. The RADseq data was weakly affected by changes of *n* parameter (for detailed results, see Tables S3.1 and S3.2). When choosing the appropriate value for *n* parameter, similar cautions should be needed as for parameter *M*. However, it is also important to note the degree of diversity among populations/individuals included in the analysis. Since increasing *n* may also increase the proportion of paralogous loci we decided to use the default value (1) as the optimal value for this parameter in our study.

The Heap tool used in this study implements a similar bioinformatics approach as GATK to call SNPs, but it seems to be less conservative [55]. In Heap, the *depth* parameter sets a threshold for the minimum number of mapped reads/nucleotides needed to call an SNP. It can be treated as an additional filtering tool for increasing confidence in returning markers that are not sequencing errors. The impact of *depth* on key indicators is a consequence of loss of markers when an individual from a replicate pair has a lower number of reads. In general, increasing *depth* distinctly decreased the number of filtered SNPs; however, GL decreased slightly or remained unchanged. In most cases, the change of GL, along with *depth*, was negatively correlated with ML (for details, see Tables S3.3 and S3.4).

The *mapq* parameter is also quality-related parameter, although it did not influence the key indicators; rather, it had an impact on the number of markers generated (for details, see Tables S3.3 and S3.4). Across the datasets, those from GBS were affected most when increasing *mapq*. Using high *mapq* values is justified, especially when the Phred quality of the delivered data is low, likely generating false-positive SNPs. Considering the 'filtering' nature of both parameters when high-quality genome and reduced representation data is available, we recommend setting them to default values (*depth* = 3; *mapq* = 20).

3.3. Whole Datasets

In our analysis, we initially decided upon the threshold that an individual would need to have at least 50% of the average number of SNP per individual to be included in subsequent analyses. In beech, all 91 samples fulfilled this criterion; however, in oak, three samples in GBS and one in RADseq did not pass the threshold, and these were discarded from further analyses.

In general, there was a congruence between the replicated samples and the whole datasets in the numbers of SNPs retained after filtering (Tables 3 and 4). The numbers of SNPs obtained based on the whole datasets depended on the genome's size, except in the case of RADseq data. The filtering process increased the sample depth from 25.12 (unfiltered) to 34.35 (filtered), which seems to be the level acceptable for most population analyses. The Ts/Tv ratio was comparable across all datasets. However, after filtering, the average Ts/Tv ratio increased by 0.18 for the de novo method and only 0.05 for the mapping method. The increased Ts/Tv ratio in the filtered sets is the expected effect of higher SNP quality after filtering [63].

Table 4. Basic statistics of results generated with de novo (Stacks) and map (HEAP) approaches on the whole datasets. (MAF—minor allele frequency; LD—linkage disequilibrium; TV/TS—transition to transversions ratio).

		Fagus sylvatica						***Quercus robur***					
		GBS		**RADseq**		**ddRAD**		**GBS**		**RADseq**		**ddRAD**	
		de novo	**map**	**de novo**	**map**	**de novo**	**map**	**de novo**	**map**	**de novo**	**map**	**de novo**	**map**
Unfiltered data	Number of samples	91	91	91	91	91	91	87	87	90	90	91	91
	Number of SNPs	124,349	58,788	157,755	130,325	391,161	482,198	325,260	180,214	257,606	304,664	406,455	634,434
	Transition	75,927	35,892	81,748	71,016	237,729	314,620	212,475	120,515	147,199	186,612	234,486	408,345
	Transversion	48,422	23,708	76,007	60,689	153,432	174,191	112,785	63,543	110,407	122,927	170,346	233,228
	TS/TV ratio	1.57	1.51	1.08	1.17	1.55	1.81	1.88	1.90	1.33	1.52	1.38	1.75
	Avg. sample depth	34.1	30.7	28.5	24.4	16.7	15.9	22.6	22.4	28.7	26.5	25.9	25.0
Filtered data SNPs > 80% + MAF >0.05 + removed LD	Number of SNPs	16,816	8270	3071	1083	28,541	28,199	28,907	15,919	709	230	35,245	36,058
	% of SNPs retained	13.52%	14.07%	1.95%	0.83%	7.30%	5.85%	8.89%	8.83%	0.28%	0.08%	8.67%	5.68%
	Transition	10,092	4979	1792	665	17,629	17,917	19,835	11,064	435	129	21,479	22,104
	Transversion	6724	3291	1279	418	10,912	10,282	9072	4855	274	101	13,766	13,954
	TS/TV ratio	1.50	1.51	1.40	1.59	1.62	1.74	2.19	2.28	1.59	1.28	1.56	1.58
	Avg. sample depth	40.5	35.4	38.2	40.8	21.9	21.4	26.3	29.4	47.7	46.4	32.8	31.4
Alternative filters with no LD filtering	SNPs >80%	39,953	24,214	11,658	2892	87,832	124,769	100,578	70,550	4033	926	132,158	172,450
	SNPs > 80% + MAF > 0.05	20,043	11,760	3707	1451	41,177	56,288	33,530	22,094	847	402	45,841	59,475
	SNPs 100%	10,237	9736	9	4	15,092	34,751	14,813	16,219	19	5	31,566	53,639
	SNPs 100% + MAF > 0.05	4769	4639	0	1	6539	16,227	4052	4704	1	1	9192	17,284

RADseq appeared to be the least efficient method for the identification and calling of SNPs. Despite the large initial number of loci, data filtering resulted in a dramatic reduction of loci that fulfilled the criteria. It seems that the most critical problem in RADseq is repeatability of loci identification. Many initial loci were filtered out due to their uniqueness (MAF > 0.05) or a high level of missing data. This correlates the high proportions of MD and MA observed in the replicated samples. We suspect that the poor RADseq performance could have resulted from low quality of input DNA samples, which may have been attributable to errors made during the construction of libraries where some samples may have not been sufficiently cleaved with a restriction enzyme or fragmented by sonicators, thus causing the dropout of many RAD sites. SNPs' final outputs based on designed filtering criteria exclude this type of library for population genomic analyses.

GBS provided reasonably large datasets for both beech and oak samples. A range of ~8000 to ~28,000 SNPs per genome (depending on species and genotyping method; Table 4) seem to be sufficient for most types of published population genomic studies; however, such density may not be adequate for in-depth association analyses. Assuming standard filtering (SNPs observed in >80% of individuals; MAF > 0.05; no LD), GBS provided more data using the de novo approach than the mapping method. This is likely related to the size of the sequence reads of GBS data. Since some shorter reads could be mapped to multiple genome positions, they were discarded as ambiguous and not considered for identification of SNPs. However, the difference in numbers of SNPs between de novo and mapping methods was minimized when SNP loci were required to be observed in all sampled individuals (SNPs: 100%), which resulted in reduced numbers of SNPs. This suggests that, for GBS, the mapping method is more conservative than the de novo approach when one allows for some level of missing data to be acceptable for downstream analyses.

The ddRAD approach appeared to be the best choice for identification and calling of SNPs in both beech and oak. After filtering, this method returned ~28,000 and ~35,000 SNPs for beech and oak, respectively. Interestingly, the locus counts were similar for de novo and mapping methods of SNP identification, which should be considered a reciprocal confirmation of the data quality. Even when requiring that a SNP locus must be genotyped among all sampled individuals (i.e., no missing data), and setting no MAF limits, there were still ~34,000 and ~53,000 markers available for beech and oak, respectively (Table 4). Such numbers of fairly repeatable SNP markers satisfy most types of population genomic studies, and should be helpful even in some genomic association analyses.

Using the available reference genomes (beech—[46]; oak—[47]), we have annotated the filtered SNPs obtained from the mapping approach to check their genomic position and usefulness in future studies. In both species, a similar pattern was observed; however, some differences between these species may result from differences in the genome annotation methods. In both cases, the GBS markers were mostly (>75%) associated with non-gene sites, while in RADseq and ddRAD a considerable proportion of SNPs was related to genes (Table 5). However, about 26% of SNPs in beech and 22% in oak were located in introns or were synonymous sites (Table 5). Considering the total number of SNPs in the filtered data sets related to genes, but excluding intron and synonymous variants, we obtained the best results with ddRAD, indicating 4105 and 3658 SNP loci in beech and oak, respectively (Table 5).

Table 5. Categories of filtered (SNPs > 80%, MAF > 0.05, LD removed) and annotated SNPs.

	Fagus sylvatica			*Quercus robur*			*Fagus sylvatica (%)*			*Quercus robur (%)*		
Annotation Type	**GBS**	**RAD**	**ddRAD**	**GBS**	**RAD**	**ddRAD**	**GBS**	**RAD**	**ddRAD**	**GBS**	**RAD**	**ddRAD**
3 prime UTR variant	165	35	687	55	1	454	2.00%	3.23%	2.44%	0.35%	0.42%	1.26%
5 prime UTR premature start codon gain variant	11	5	49	-	-	43	0.13%	0.46%	0.17%	-	-	0.12%
5 prime UTR variant	23	14	242	5	-	260	0.28%	1.29%	0.86%	0.03%	-	0.72%
Initiator codon variant	-	-	3	-	-	-	-	-	0.01%	-	-	-
Missense variant	298	205	2690	180	27	2521	3.60%	18.93%	9.54%	1.13%	11.44%	6.99%
Missense variant&splice region variant	9	3	41	7	-	40	0.11%	0.28%	0.15%	0.04%	-	0.11%
Splice acceptor variant&intron variant	3	1	27	1	-	4	0.04%	0.09%	0.10%	0.01%	-	0.01%
Splice donor variant&intron variant	3	-	15	1	-	4	0.04%	-	0.05%	0.01%	-	0.01%
Splice region variant	3	1	12	-	-	5	0.04%	0.09%	0.04%	-	-	0.01%
Splice region variant&intron variant	37	13	177	21	-	243	0.45%	1.20%	0.63%	0.13%	-	0.67%
Splice region variant&synonymous variant	6	1	31	4	-	37	0.07%	0.09%	0.11%	0.03%	-	0.10%
Start lost	-	-	4	1	-	4	-	-	0.01%	0.01%	-	0.01%
Stop gained	8	7	114	8	1	36	0.10%	0.65%	0.40%	0.05%	0.42%	0.10%
Stop gained&splice region variant	-	-	4	-	-	1	-	-	0.01%	-	-	0.00%
Stop lost	1	-	8	-	-	1	0.01%	-	0.03%	-	-	0.00%
Stop lost&splice region variant	-	-	-	-	-	2	-	-	-	-	-	0.01%
Stop retained variant	1	-	1	-	-	3	0.01%	-	0.00%	-	-	0.01%
Synonymous variant	253	201	1963	145	25	2653	3.06%	18.56%	6.96%	0.91%	10.59%	7.36%
Intron variant	1204	125	3418	1234	27	6066	14.56%	11.54%	12.12%	7.75%	11.44%	16.82%
Intergenic region	830	73	2597	10,015	83	12,212	10.04%	6.74%	9.21%	62.91%	35.17%	33.87%
Downstream gene variant	2033	178	6249	2046	37	6343	24.58%	16.44%	22.16%	12.85%	15.68%	17.59%
Upstream gene variant	3382	221	9867	2196	35	5126	40.89%	20.41%	34.99%	13.79%	14.83%	14.22%
Total	8270	1083	28,199	15,919	236	36,058						
Non genes	6245	472	18,713	14,257	155	23,681	75.51%	43.58%	66.36%	89.56%	65.68%	65.67%
Genes	2025	611	9486	1662	81	12377	24.49%	56.42%	33.64%	10.44%	34.32%	34.33%
Genes—excluding intron and synonymous	568	285	4105	283	29	3658	6.87%	26.32%	14.56%	1.78%	12.29%	10.14%

4. Discussion

In recent years, a growing interest in reduced-representation genomic approaches and their applications proved their usefulness in numerous studies [8,64,65]. However, the variety of available library types and analytical pipelines to process this data can be confusing, especially for researchers with limited experience, despite the broad support of research community. The use of technical replicates in the optimization of SNP identification and calling is described in many studies [19,66–68] and it appears to be a good practice to tune the pipeline parameters to the species and scope of the study, and to monitor the quality of SNP identification. Our results confirm these findings and provide additional information in the discussion on how different types of libraries from the same species can be influenced by distinct genotyping strategies.

Testing of technical replicates is inexpensive, and it should be the initial requirement when using RAD-based genomic libraries for marker discovery. The optimization step can provide initial insights into the expected range of results in a larger study, and reduce risk of project failure. The analytical process after data delivery is fast, and evaluation of a few samples on a standard computer (e.g., four-thread CPU with 8 GB of RAM) takes from 24 to 72 h. Due to the rapid development of out-of-the-box solutions, both offline (e.g., dDocent [61]) and online (e.g., Galaxy [69]), the data analysis process can be performed by staff even only moderately involved in the field of bioinformatics. However, to obtain informative SNPs, specific requirements for the isolated DNA must be met [70].

NGS is particularly sensitive to low DNA quality, and vulnerable to errors which can emerge during the library preparation [71,72]. Our results suggest that laboratory errors can lead to genomic region/marker dropout, as in the case of our RADseq dataset, despite a satisfying number of reads per sample. However, RADseq is known to suffer from large proportions of missing data [73]. When outsourcing the construction of libraries and subsequent sequencing, research teams should be focused on providing good quality DNA or fresh raw material to avoid potential data errors [74].

Overrepresentation in de novo SNP data, as observed in GBS datasets, can result from differences in genome size among individuals within the species [75,76], especially when using a restriction enzyme that cuts frequently, as illustrated in our GBS data. When the mapping approach is used, sequencing errors may decrease SNP numbers after unfitting bases are discarded. Nonetheless, this reference-based strategy would provide data that may be more comparable with that of other studies, and it can also deliver additional information e.g., annotation of SNPs [8]. This genotyping strategy should always be the first choice when a reference genome is available [77]. On the other hand, although the availability of plant genomes is increasing on a daily basis, they are still scarce. In the absence of a reference genome, de novo genotyping is a reliable alternative, as demonstrated in other studies [78].

It should be noted that in this study, we used an initial filtering of SNP data (SNPs biallelic; present in 75–80% of samples; indels removed), which generally alleviates some common genotyping problems, including missing loci. In particular, the choice to filter out those loci present in less than a specified fraction of samples (e.g., 80%) appears to be an efficient way of pre-selecting reliable loci [8,79]. Therefore, our variation of Stacks or Heap parameters had only a moderate effect on the size and quality of resulting SNP datasets. After the initial optimization, based on replicated samples, the choice of the correct parameters to genotype the whole sample set is always the trickiest part of the study. Here, we would not like to provide unequivocal information on what exact values should be chosen, because these depend on the scope of analysis, the species' biology, and even sampling strategy (local or wide sampling). However, we intend to share some observations and guidelines on how to perform the optimization process.

- The assessment of the raw and filtered numbers of markers (i.e., biallelic SNPs, present in at least 6 out of 8 samples, indels removed) helps to detect library/data errors. For example, significant loss of markers in a filtered set will usually be a signal of uneven genome coverage by reads, regardless of the genotyping method.

- When assessing the number of markers reported based on parameters associated with more restrictive genotyping (e.g., *m*, *depth*, *mapq*), first check the number of SNPs after filtering (with the abovementioned criteria). In some sets, more stringent filtering did not cause the expected improvement but rather the possible loss of SNPs.
- Pairs of replicated samples from the same individuals will always share a significantly higher proportion of good loci (GL) with each other than with any other sample. This observation can be used as a quality control tool to determine whether a swap of samples occurred.
- To overcome a threat of false positive SNPs, which can occur with highly elevated numbers of markers, we suggest focusing on the proportion of good loci as the most important key indicator.
- Increasing the minimum number of reads necessary to create a stack (*m*) will always decrease read number and cause dropout of underrepresented makers leading to the decreased levels of GL and a higher proportion of ML (as in the case of GBS and ddRAD) and, in cases of problems with data uniformity throughout the whole genome, shifted MA values (as in the case of RADseq).
- If a reference genome for the species under analysis is available, optimization should be conducted using the reference-based approach. It can be expected to deliver results that are more reliable, and more comparable to those of other studies.
- The influence of SNP-calling procedure has a profound effect on the number and quality of markers. Both *depth* and *mapq* should be treated as filtering tools; high values for these parameters will significantly decrease the number of markers returned. If more stringent filtering is necessary, using elevated *mapq* is a preferable option due to having no effect on key indicators. These symptoms are usually a sign of increased detection of false positive SNPs [19].
- Shifted mismatch both on an individual level and in the joint catalog should be adjusted with respect to species biology and the sampling strategy applied [66].

5. Conclusions

Genotyping by next-generation sequencing (NGS) of reduced-representation genomic libraries associated with restriction enzymes became a common approach to identify large numbers of genetic markers (mostly SNPs) uniformly distributed across genomes. However, the number and quality of RAD-based markers obtained in particular studies depends on many aspects, including the quality of DNA isolation, the choice of RRL, the type of restriction enzymes, the design of sequencing (resulting in sequencing depth), and the bioinformatics pipelines used for identification and calling of SNPs [8,22]. Testing all of the possible variables is beyond the scope of a single study. In this paper, we briefly evaluated the three RRL approaches (GBS, RADseq, ddRAD) and different methods of SNP identification (de novo or by reference genome mapping) to find the best toolset for future population genomics studies in two broadleaved tree species: *F. sylvatica* and *Q. robur*.

We found that the most promising approach—providing relatively large numbers of reliable SNPs—is to employ the ddRAD technique and a calling approach based on mapping sequence reads to a reference genome. Based on about 90 individuals within species, we found ~28,000 and ~36,000 loci for beech and oak, respectively, given typical filtering criteria (MAF > 0.05; SNPs present in >80% samples; LD $r^2 < 0.5$). However, when relaxing LD filtering limitations, these numbers increased up to ~56,000 and ~59,000 respectively (Table 4). Based on technical replicates, we estimated that in ddRAD more than 80% of SNP loci should be considered reliable. Additionally, according to annotations on the reference genomes, we found that in both species more than 30% of the identified loci could be related to genes. These findings provide a solid support for the use of ddRAD-based SNPs for future population genomics, or even for genomics selection studies.

Supplementary Materials: The following are available online at https://www.mdpi.com/1999-4907/12/2/222/s1, DNA isolation protocol, Table S1.1.—European beech (*Fagus sylvatica* L.) sample basic information; Table S1.2.—Pedunculate oak (*Quercus robur* L.) sample basic information; Table S2. Basic

replicate samples information; Table S3.1. Results of parameters (*m*, *M*, *n*) altering for genotyping with stacks (de novo method) of European beech (*Fagus sylvatica* L.) replicated samples; Table S3.2. Results of parameters (*m*, *M*, *n*) altering for genotyping with stacks (de novo method) of pedunculate oak (*Quercus robur* L.) replicated samples; Table S3.3. Results of parameters (*depth*, *mapq*) altering for genotyping with Heap (mapping method) of European beech (*Fagus sylvatica* L.) replicated samples; Table S3.4. Results of parameters (*depth*, *mapq*) altering for genotyping with Heap (mapping method) of pedunculate oak (*Quercus robur* L.) replicated samples.

Author Contributions: Conceptualization, B.U. and J.B.; Methodology, B.U. and J.M.; Investigation, B.U. and J.M.; Data curation, B.U.; Writing—original draft preparation, B.U.; Writing—review and editing, B.U., J.M., and J.B.; Supervision, J.B.; Project administration, J.B.; Funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: The study was supported by the National Science Center, Poland (2012/04/A/NZ9/00500), and the Polish Ministry of Science and Higher Education under the program "Regional Initiative of Excellence" in 2019–2022 (grant no. 008/RID/2018/19).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw GBS, RADseq and ddRAD data have been submitted to NCBI SRA database under the BioProject ID number PRJNA694960.

Acknowledgments: We would like to thank Władysław Barzdajn from Poznań University of Life Sciences and the staff of Experimental Forest District in Siemianice as well as Roman Rożkowski from the Institute of Dendrology, Polish Academy of Science and the staff of the Choszczno Forest District for their support on the study sites. We also to thank our lab team members: Ewa Sztupecka and Katarzyna Meyza for their outstanding job in fieldwork and DNA isolations.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Straub, S.C.; Parks, M.; Weitemier, K.; Fishbein, M.; Cronn, R.C.; Liston, A. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Am. J. Bot.* **2012**, *99*, 349–364. [CrossRef]
2. Unamba, C.I.N.; Nag, A.; Sharma, R.K. Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. *Front. Plant Sci.* **2015**, 6. [CrossRef]
3. Kim, C.; Guo, H.; Kong, W.; Chandnani, R.; Shuang, L.-S.; Paterson, A.H. Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci.* **2016**, *242*, 14–22. [CrossRef]
4. Wang, R.; Fan, J.; Chang, P.; Zhu, L.; Zhao, M.; Li, L. Genome Survey Sequencing of *Acer truncatum* Bunge to Identify Genomic Information, Simple Sequence Repeat (SSR) Markers and Complete Chloroplast Genome. *Forests* **2019**, *10*, 87. [CrossRef]
5. Fuentes-Pardo, A.P.; Ruzzante, D.E. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol. Ecol.* **2017**, *26*, 5369–5406. [CrossRef]
6. Altshuler, D.; Pollara, V.J.; Cowles, C.R.; Van Etten, W.J.; Baldwin, J.; Linton, L.; Lander, E.S. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **2000**, *407*, 513–516. [CrossRef]
7. Okou, D.T.; Steinberg, K.M.; Middle, C.; Cutler, D.J.; Albert, T.J.; Zwick, M.E. Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **2007**, *4*, 907–909. [CrossRef]
8. Andrews, K.R.; Good, J.M.; Miller, M.R.; Luikart, G.; Hohenlohe, P.A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **2016**, *17*, 81–92. [CrossRef]
9. Baird, N.A.; Etter, P.D.; Atwood, T.S.; Currey, M.C.; Shiver, A.L.; Lewis, Z.A.; Selker, E.U.; Cresko, W.A.; Johnson, E.A. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **2008**, *3*, e3376. [CrossRef]
10. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **2011**, *6*, e19379. [CrossRef]
11. Peterson, B.K.; Weber, J.N.; Kay, E.H.; Fisher, H.S.; Hoekstra, H.E. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **2012**, *7*, e37135. [CrossRef]
12. Hohenlohe, P.A.; Hand, B.K.; Andrews, K.R.; Luikart, G. Population genomics provides key insights in ecology and evolution. In *Population Genomics*; Springer: Cham, Switzerland, 2018; pp. 483–510.
13. O'Leary, S.J.; Puritz, J.B.; Willis, S.C.; Hollenbeck, C.M.; Portnoy, D.S. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* **2018**. [CrossRef]
14. Annicchiarico, P.; Nazzicari, N.; Pecetti, L.; Romani, M.; Ferrari, B.; Wei, Y.; Brummer, E.C. GBS-Based Genomic Selection for Pea Grain Yield under Severe Terminal Drought. *Plant Genome* **2017**, *10*. [CrossRef]

15. Davey, J.W.; Cezard, T.; Fuentes-Utrilla, P.; Eland, C.; Gharbi, K.; Blaxter, M.L. Special features of RAD Sequencing data: Implications for genotyping. *Mol. Ecol.* **2013**, *22*, 3151–3164. [CrossRef]
16. Arnold, B.; Corbett-Detig, R.B.; Hartl, D.; Bomblies, K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **2013**, *22*, 3179–3190. [CrossRef]
17. Li, H.; Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **2010**, *11*, 473–483. [CrossRef] [PubMed]
18. Ruffalo, M.; LaFramboise, T.; Koyuturk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **2011**, *27*, 2790–2796. [CrossRef]
19. Mastretta-Yanes, A.; Arrigo, N.; Alvarez, N.; Jorgensen, T.H.; Pinero, D.; Emerson, B.C. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* **2015**, *15*, 28–41. [CrossRef] [PubMed]
20. Mayer-Jochimsen, M.; Fast, S.; Tintle, N.L. Assessing the Impact of Differential Genotyping Errors on Rare Variant Tests of Association. *PLoS ONE* **2013**, *8*, e56626. [CrossRef] [PubMed]
21. Nielsen, R.; Paul, J.S.; Albrechtsen, A.; Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **2011**, *12*, 443–451. [CrossRef] [PubMed]
22. Parchman, T.L.; Jahner, J.P.; Uckele, K.A.; Galland, L.M.; Eckert, A.J. RADseq approaches and applications for forest tree genetics. *Tree Genet. Genomes* **2018**, *14*, 39. [CrossRef]
23. Fountain, E.D.; Pauli, J.N.; Reid, B.N.; Palsbøll, P.J.; Peery, M.Z. Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Mol. Ecol. Resour.* **2016**, *16*, 966–978. [CrossRef] [PubMed]
24. Maroso, F.; Hillen, J.E.J.; Pardo, B.G.; Gkagkavouzis, K.; Coscia, I.; Hermida, M.; Franch, R.; Hellemans, B.; Van Houdt, J.; Simionati, B.; et al. Performance and precision of double digestion RAD (ddRAD) genotyping in large multiplexed datasets of marine fish species. *Mar. Genom.* **2018**, *39*, 64–72. [CrossRef]
25. Shafer, A.B.A.; Peart, C.R.; Tusso, S.; Maayan, I.; Brelsford, A.; Wheat, C.W.; Wolf, J.B.W. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods Ecol. Evol.* **2017**, *8*, 907–917. [CrossRef]
26. Catchen, J.; Hohenlohe, P.A.; Bassham, S.; Amores, A.; Cresko, W.A. Stacks: An analysis tool set for population genomics. *Mol. Ecol.* **2013**, *22*, 3124–3140. [CrossRef]
27. Tuskan, G.A.; Difazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A.; et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **2006**, *313*, 1596–1604. [CrossRef]
28. Myburg, A.A.; Grattapaglia, D.; Tuskan, G.A.; Hellsten, U.; Hayes, R.D.; Grimwood, J.; Jenkins, J.; Lindquist, E.; Tice, H.; Bauer, D.; et al. The genome of *Eucalyptus grandis*. *Nature* **2014**, *510*, 356–362. [CrossRef]
29. Zimin, A.; Stevens, K.A.; Crepeau, M.W.; Holtz-Morris, A.; Koriabine, M.; Marcais, G.; Puiu, D.; Roberts, M.; Wegrzyn, J.L.; de Jong, P.J.; et al. Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics* **2014**, *196*, 875–890. [CrossRef]
30. Cruz, F.; Julca, I.; Gómez-Garrido, J.; Loska, D.; Marcet-Houben, M.; Cano, E.; Galán, B.; Frias, L.; Ribeca, P.; Derdak, S.; et al. Genome sequence of the olive tree, *Olea europaea*. *GigaScience* **2016**, 5. [CrossRef]
31. Sork, V.L.; Fitz-Gibbon, S.T.; Puiu, D.; Crepeau, M.; Gugger, P.F.; Sherman, R.; Stevens, K.; Langley, C.H.; Pellegrini, M.; Salzberg, S.L. First Draft Assembly and Annotation of the Genome of a California Endemic Oak *Quercus lobata* Nee (Fagaceae). *G3* **2016**. [CrossRef]
32. Petit, R.J.; Hampe, A. Some Evolutionary Consequences of Being a Tree. *Annu. Rev. Ecol. Evol. Syst.* **2006**, *37*, 187–214. [CrossRef]
33. Hamrick, J.L.; Godt, M.; Sherman-Broyles, S. Factors influencing levels of genetic diversity in woody plant species. In *Population Genetics of Forest Trees*; Adams, W.T., Strauss, S., Copes, D., Griffin, A.R., Eds.; Springer: Amsterdam, The Netherlands, 1992; Volume 42, pp. 95–124.
34. Merzeau, D.; Comps, B.; Thiébaut, B.; Letouzey, J. Estimation of *Fagus sylvatica* L mating system parameters in natural populations. *Ann. Sci.* **1994**, *51*, 163–173. [CrossRef]
35. Chybicki, I.J.; Burczyk, J. Seeing the forest through the trees: Comprehensive inference on individual mating patterns in a mixed stand of *Quercus robur* and *Q. petraea*. *Ann. Bot.* **2013**, *112*, 561–574. [CrossRef]
36. Barbier, S.; Gosselin, F.; Balandier, P. Influence of tree species on understory vegetation diversity and mechanisms involved—A critical review for temperate and boreal forests. *For. Ecol. Manag.* **2008**, *254*, 1–15. [CrossRef]
37. Packham, J.R.; Thomas, P.A.; Atkinson, M.D.; Degen, T. Biological Flora of the British Isles: *Fagus sylvatica*. *J. Ecol.* **2012**, *100*, 1557–1608. [CrossRef]
38. Eaton, E.; Caudullo, G.; Oliveira, S.; De Rigo, D. Quercus robur and Quercus petraea in Europe: Distribution, habitat, usage and threats. In *European Atlas of Forest Tree Species*; San-Miguel-Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T., Mauri, A., Eds.; Publications Office of the European Union: Luxembourg, 2016; pp. e01c6df, 160–163.
39. Cuervo-Alarcon, L.; Arend, M.; Muller, M.; Sperisen, C.; Finkeldey, R.; Krutovsky, K.V. Genetic variation and signatures of natural selection in populations of European beech (*Fagus sylvatica* L.) along precipitation gradients. *Tree Genet. Genomes* **2018**, *14*, 84. [CrossRef]
40. Caignard, T.; Delzon, S.; Bodenes, C.; Dencausse, B.; Kremer, A. Heritability and genetic architecture of reproduction-related traits in a temperate oak species. *Tree Genet. Genomes* **2019**, *15*, 1. [CrossRef]
41. Müller, M.; Seifert, S.; Finkeldey, R. A candidate gene-based association study reveals SNPs significantly associated with bud burst in European beech (*Fagus sylvatica* L.). *Tree Genet. Genomes* **2015**, *11*, 1–13. [CrossRef]

42. Pluess, A.R.; Frank, A.; Heiri, C.; Lalague, H.; Vendramin, G.G.; Oddou-Muratorio, S. Genome-environment association study suggests local adaptation to climate at the regional scale in *Fagus sylvatica*. *New Phytol.* **2016**, *210*, 589–601. [CrossRef]
43. Krajmerová, D.; Hrivnák, M.; Ditmarová, Ľ.; Jamnická, G.; Kmeť, J.; Kurjak, D.; Gömöry, D. Nucleotide polymorphisms associated with climate, phenology and physiological traits in European beech (*Fagus sylvatica* L.). *New For.* **2017**, *48*, 463–477. [CrossRef]
44. Hipp, A.L.; Manos, P.S.; Hahn, M.; Avishai, M.; Bodenes, C.; Cavender-Bares, J.; Crowl, A.A.; Deng, M.; Denk, T.; Fitz-Gibbon, S.; et al. Genomic landscape of the global oak phylogeny. *New Phytol.* **2019**. [CrossRef]
45. Meger, J.; Ulaszewski, B.; Vendramin, G.G.; Burczyk, J. Using reduced representation libraries sequencing methods to identify cpDNA polymorphisms in European beech (*Fagus sylvatica* L). *Tree Genet. Genomes* **2019**, *15*, 7. [CrossRef]
46. Mishra, B.; Gupta, D.K.; Pfenninger, M.; Hickler, T.; Langer, E.; Nam, B.; Paule, J.; Sharma, R.; Ulaszewski, B.; Warmbier, J.; et al. A reference genome of the European beech (*Fagus sylvatica* L.). *Gigascience* **2018**, *7*, giy063. [CrossRef]
47. Plomion, C.; Aury, J.M.; Amselem, J.; Leroy, T.; Murat, F.; Duplessis, S.; Faye, S.; Francillonne, N.; Labadie, K.; Le Provost, G.; et al. Oak genome reveals facets of long lifespan. *Nat. Plants* **2018**, *4*, 440–452. [CrossRef]
48. Barzdajn, W. Proweniencyjna zmienność buka zwyczajnego [*Fagus sylvatica* L.] w Polsce w świetle wyników doświadczenia proweniencyjnego serii 1992/1995. *Sylwan* **2002**, *146*, 5–34.
49. Chmura, D.J.; Guzicka, M.; Rożkowski, R.; Michałowicz, D.; Grodzicki, W.; Chałupka, W. Produktywność biomasy nadziemnej i podziemnej w doświadczeniu proweniencyjno—rodowym z dębem szypułkowym. *Sylwan* **2014**, *158*, 829–839.
50. Healey, A.; Furtado, A.; Cooper, T.; Henry, R.J. Protocol: A simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **2014**, *10*, 21. [CrossRef] [PubMed]
51. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 16 November 2020).
52. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef]
53. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
54. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595. [CrossRef] [PubMed]
55. Kobayashi, M.; Ohyanagi, H.; Takanashi, H.; Asano, S.; Kudo, T.; Kajiya-Kanegae, H.; Nagano, A.J.; Tainaka, H.; Tokunaga, T.; Sazuka, T.; et al. Heap: A highly sensitive and accurate SNP detection tool for low-coverage high-throughput sequencing data. *DNA Res.* **2017**, *24*, 397–405. [CrossRef]
56. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef]
57. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]
58. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Subgroup, G.P.D.P. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
59. Team, R.C. R: A Language and Environment for Statistical Computing. Available online: https://www.r-project.org (accessed on 24 January 2020).
60. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **2012**, *6*, 80–92. [CrossRef]
61. Puritz, J.B.; Hollenbeck, C.M.; Gold, J.R. dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2014**, *2*, e431. [CrossRef]
62. Kececioglu, J.; DeBlasio, D. Accuracy estimation and parameter advising for protein multiple sequence alignment. *J. Comput. Biol.* **2013**, *20*, 259–279.
63. Wang, J.; Raskin, L.; Samuels, D.C.; Shyr, Y.; Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* **2014**, *31*, 318–323. [CrossRef]
64. Nagamitsu, T.; Uchiyama, K.; Izuno, A.; Shimizu, H.; Nakanishi, A. Environment-dependent introgression from *Quercus dentata* to a coastal ecotype of *Quercus mongolica* var. *crispula* in northern Japan. *New Phytol.* **2020**, *226*, 1018–1028. [CrossRef]
65. Schley, R.J.; Pennington, R.T.; Pérez-Escobar, O.A.; Helmstetter, A.J.; de la Estrella, M.; Larridon, I.; Sabino Kikuchi, I.A.B.; Barraclough, T.G.; Forest, F.; Klitgård, B. Introgression across evolutionary scales suggests reticulation contributes to Amazonian tree diversity. *Mol. Ecol.* **2020**, *29*, 4170–4185. [CrossRef] [PubMed]
66. Aguirre, N.C.; Filippi, C.V.; Zaina, G.; Rivas, J.G.; Acuña, C.V.; Villalba, P.V.; García, M.N.; González, S.; Rivarola, M.; Martínez, M.C.; et al. Optimizing ddRADseq in Non-Model Species: A Case Study in *Eucalyptus dunnii* Maiden. *Agronomy* **2019**, *9*, 484. [CrossRef]
67. McCartney-Melstad, E.; Gidiş, M.; Shaffer, H.B. An empirical pipeline for choosing the optimal clustering threshold in RADseq studies. *Mol. Ecol. Resour.* **2019**, *19*, 1195–1204. [CrossRef]
68. Bresadola, L.; Link, V.; Buerkle, C.A.; Lexer, C.; Wegmann, D. Estimating and accounting for genotyping errors in RAD-seq experiments. *Mol. Ecol. Resour.* **2020**. [CrossRef]

69. Giardine, B.; Riemer, C.; Hardison, R.C.; Burhans, R.; Elnitski, L.; Shah, P.; Zhang, Y.; Blankenberg, D.; Albert, I.; Taylor, J.; et al. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* **2005**, *15*, 1451–1455. [CrossRef]
70. Peterson, G.W.; Dong, Y.; Horbach, C.; Fu, Y.-B. Genotyping-By-Sequencing for Plant Genetic Diversity Analysis: A Lab Guide for SNP Genotyping. *Diversity* **2014**, *6*, 665–680. [CrossRef]
71. Cumer, T.; Pouchon, C.; Boyer, F.; Yannic, G.; Rioux, D.; Bonin, A.; Capblancq, T. Double-digest RAD-sequencing: Do wet and dry protocol parameters impact biological results? *bioRxiv* **2018**. [CrossRef]
72. Graham, C.F.; Glenn, T.C.; McArthur, A.G.; Boreham, D.R.; Kieran, T.; Lance, S.; Manzon, R.G.; Martino, J.A.; Pierson, T.; Rogers, S.M.; et al. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Mol. Ecol. Resour.* **2015**, *15*, 1304–1315. [CrossRef]
73. Tripp, E.A.; Tsai, Y.-H.E.; Zhuang, Y.; Dexter, K.G. RADseq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceae) in the ultra-arid deserts of Namibia. *Ecol. Evol.* **2017**, *7*, 7920–7936. [CrossRef]
74. Touchman, J.W.; Mastrian, S.D. DNA Sequencing: An Outsourcing Guide. *Curr. Protoc. Essent. Lab. Tech.* **2008**. [CrossRef]
75. Biémont, C. Genome size evolution: Within-species variation in genome size. *Heredity* **2008**, *101*, 297–298. [CrossRef] [PubMed]
76. Voronova, A.; Belevich, V.; Korica, A.; Rungis, D. Retrotransposon distribution and copy number variation in gymnosperm genomes. *Tree Genet. Genomes* **2017**, *13*, 88. [CrossRef]
77. Torkamaneh, D.; Laroche, J.; Belzile, F. Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. *PLoS ONE* **2016**, *11*, e0161333. [CrossRef]
78. Fitz-Gibbon, S.; Hipp, A.L.; Pham, K.K.; Manos, P.S.; Sork, V.L. Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* section *Quercus*). *Genome* **2017**, *60*, 743–755. [CrossRef] [PubMed]
79. Gargiulo, R.; Kull, T.; Fay, M.F. Effective double-digest RAD sequencing and genotyping despite large genome size. *Mol. Ecol. Resour.* **2020**. [CrossRef]