forests

MDPI

Article

Comparing the Effectiveness of Exome Capture Probes, Genotyping by Sequencing and Whole-Genome Re-Sequencing for Assessing Genetic Diversity in Natural and Managed Stands of *Picea abies*

Helena Eklöf [1], Carolina Bernhardsson [2] and Pär K. Ingvarsson [3,*]

[1] Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, SE901 87 Umeå, Sweden; helena.eklof@umu.se
[2] Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, SE752 36 Uppsala, Sweden; carolina.bernhardsson@ebc.uu.se
[3] Department of Plant Biology, Linnean Centre for Plant Biology, Swedish University of Agricultural Sciences, SE750 07 Uppsala, Sweden
* Correspondence: par.ingvarsson@slu.se

Received: 8 October 2020; Accepted: 5 November 2020; Published: 10 November 2020

Abstract: Conifer genomes are characterized by their large size and high abundance of repetitive material, making large-scale genotyping in conifers complicated and expensive. One of the consequences of this is that it has been difficult to generate data on genome-wide levels of genetic variation. To date, researchers have mainly employed various complexity reduction techniques to assess genetic variation across the genome in different conifer species. These methods tend to capture variation in a relatively small subset of a typical conifer genome and it is currently not clear how representative such results are. Here we take advantage of data generated in the first large-scale re-sequencing effort in Norway spruce and assess how well two commonly used complexity reduction methods, targeted capture probes and genotyping by sequencing perform in capturing genome-wide variation in Norway spruce. Our results suggest that both methods perform reasonably well for assessing genetic diversity and population structure in Norway spruce (*Picea abies* (L.) H. Karst.). Targeted capture probes were slightly more effective than GBS, likely due to them targeting known genomic regions whereas the GBS data contains a substantially greater fraction of repetitive regions, which sometimes can be problematic for assessing genetic diversity. In conclusion, both methods are useful for genotyping large numbers of samples and they greatly reduce the cost involved with genotyping a species with such a complex genome as Norway spruce.

Keywords: genetic diversity; genotyping; GBS; Norway spruce; *Picea abies*

1. Introduction

Norway spruce (*Picea abies* (L.) Karst.) is an evergreen gymnosperm tree that belongs to the biggest group of gymnosperms, the family Pinaceae [1]. It is one of the most important conifer species in Europe, with a distribution range extending from the west coast of Norway to mainland Russia in northern Europe and across the Alps, the Carpathians and the Balkans in central Europe [2]. Norway spruce is a wind pollinated and predominantly outcrossing [3] and this creates high levels of gene flow that is one of the most important factors influencing the genetic structure of populations. High gene flow results in high genetic diversity within populations and reduces the difference between populations [3]. Conifers, as a group, are characterized by high levels of pollen gene flow and this is

often found in studies of seed orchards contamination from surrounding forests [4,5]. A review study completed on six different conifer species gave an average contamination rate of seed orchards of 45% [6]. In Norway spruce seed orchards, high levels of pollen contamination have also been recorded. In Finland a seed orchard was found to have pollen contamination rates of 69–71% [7] and a study on two seed orchards in Sweden reported 43% and 59% of pollen contamination, respectively [8,9]. Scotti et al. [10] showed that seed dispersal creates patches of genotype diversity in mitochondrial markers that arise due to the limited long-distance dispersal seen in seeds. Pollen dispersal (based on chloroplast markers), on the other hand, had a homogenizing effect over both short and long distances due to the effective wind dispersal of pollen.

Pinaceae genomes have a largely conserved karyotype, consisting of 12 ($2n$ = 24) chromosomes [11–13] and in *Picea* these are of similar size and shape [12,14]. Most conifer genomes are also characterized by their large size (>15 Gbp) and high abundance of repetitive material, making sequencing and assembling a draft genome a daunting task in conifers. Nystedt et al. [14] published the first draft assembly for Norway spruce and found little evidence that recent whole genome duplications are responsible for the large genome size of *Picea abies* and other conifers. The genome contains a high abundance of repetitive sequences (around 60%) belonging to different families of transposable elements that have expanded and inserted into the spruce genome over tens of millions years, creating large introns and a high number of pseudogenes [14]. One consequence of the large and complex genomes of conifers is that it has proven difficult to generate data on genome-wide levels of genetic variation. Early studies aimed at characterizing genetic diversity in conifers were mainly focused on assessing variation in short genic [15,16] or non-coding regions [17,18]. To be able to get a more unbiased view of genome-wide variation in Norway spruce, researchers have recently employed various complexity reduction techniques, such as genotyping by sequencing (GBS), restriction digest associated sequencing (RADseq) or targeted capture probes [19–22]. However, these methods also capture variation in a relatively small subset of a typical conifer genome and it is not clear how representative such results are.

Even with current sequencing technologies, resequencing a single conifer genome is expensive and time consuming and this effectively limits the number of individuals that can be assessed. The large and repetitive nature of a typical conifer genome also make handling of re-sequencing data difficult, due to, for instance, missing regions in the genome assembly and multi-mapping of reads derived from repetitive regions. As the first studies employing whole-genome re-sequencing have been performed in conifers [23,24] it is therefore interesting to assess how well different complexity reduction methods capture genome-wide variation in genetic diversity in a typical conifer genome. Complexity reductions methods are used to capture a smaller fraction of the target genome, and this is usually achieved by either selecting specific regions using digestions with restriction enzymes (e.g., RADseq and GBS) or by using capture probes targeting specific genomic regions. In this paper we take advantage of data generated in the first large-scale re-sequencing effort in Norway spruce [24] and use this to assess how well two commonly used complexity reduction methods, targeted capture probes [22] and genotyping by sequencing [19,21], perform in capturing genome-wide variation in Norway spruce.

2. Method

2.1. Sampling

We sampled a total of 34 trees from the natural distribution range of Norway spruce (*Picea abies*). Individuals were collected from different populations in Finland, Sweden, Norway, Poland, Belarus and Romania (locations for all samples is shown in Table 1). Samples were taken either as dormant buds or fresh needles and stored in −80 °C until DNA extraction. The 20 individuals that were sampled in northern Sweden (Västerbotten county) were sampled from either young and planted (<20 years old, Marsfjället 12.3 and Långrumpskogen 2.3) or old and untouched (>150 years old, Marsfjället 12.1 and Långrumpskogen 2.1) stands in the same general location. Marsfjället is located close to the

Norwegian border in the mountain area in the west of Sweden where as Långrumpskogen is located close to the eastern coast of Sweden.

Table 1. Location for all samples.

Sample	Location	Population
Pab002	Gettinge, Sweden	-
Pab003	Vitebsk, Belarus	-
Pab004	Blizyn, Poland	-
Pab005	Toplita, Romania	-
Pab006	Köttsjön, Sweden	-
Pab007	Hatfjelldal, Norway	-
Pab008	Rovaniemen, Finland	-
Pab009	Kittilä, Finland	-
Pab010	Suomussalmi, Finland	-
Pab011	Hemnes, Norway	-
Pab012	Levanger, Norway	-
Pab013	Grane, Norway	-
Pab014	Tyda, Norway	-
Pab015	Loppi, Sweden	-
Pab016	Marsfjället Planted	12.3
Pab017	Marsfjället Planted	12.3
Pab018	Marsfjället Planted	12.3
Pab019	Marsfjället Planted	12.3
Pab020	Marsfjället Planted	12.3
Pab021	Marsfjället Old	12.1
Pab022	Marsfjället Old	12.1
Pab023	Marsfjället Old	12.1
Pab024	Marsfjället Old	12.1
Pab025	Marsfjället Old	12.1
Pab026	Långrumpskogen Planted	2.3
Pab027	Långrumpskogen Planted	2.3
Pab028	Långrumpskogen Planted	2.3
Pab029	Långrumpskogen Planted	2.3
Pab030	Långrumpskogen Planted	2.3
Pab031	Långrumpskogen Old	2.1
Pab032	Långrumpskogen Old	2.1
Pab033	Långrumpskogen Old	2.1
Pab034	Långrumpskogen Old	2.1
Pab035	Långrumpskogen Old	2.1

2.2. Whole-Genome Re-Sequencing Data

The sequencing and SNP calling procedures for these samples have previously been described in detail in Bernhardsson et al. [25] and Wang et al. [24]. Briefly, DNA was extracted with a Qiagen plant mini kit (Qiagen, Hilden, Germany) according to the instructions for the WGS data. Sequencing was performed at the National Genomics Initiative platform (NGI) at SciLifeLab in Stockholm, Sweden. Paired-end libraries with an insert size of 500 bp were sequenced using either the Illumina HiSeq 2000 (Pab002–Pab006) or Illumina HiSeq X platforms (Pab007–Pab034). Samples sequenced on both platforms were analysed using the same general bioinformatics pipeline [25]. Raw sequence reads

were mapped against the *P. abies* reference genome v.1.0 [14] using BWA-MEM with default settings. Following read mapping against the entire reference genome, the reference genome was reduced by only keeping genomic scaffolds greater than 1kb in length [26]. Before SNP calling, PCR duplicates, generated during PCR steps in sequencing library construction, were marked in all data subsets using MarkDuplicates in Picard v2.0.1 (http://broadinstitute.github.io/picard/) to eliminate artefacts introduced due to DNA amplification by polymerase chain reaction (PCR). Artefacts in read alignments occurring in regions with insertions and/or deletions (indels) were addressed using local realignment with GATK v3.7 [27]. Finally, SNP calling was performed using GATK HaplotypeCaller. Variant filtering was performed to only include biallelic SNPs that were (i) positioned >5 bp away from an indel, (ii) fulfilled GATKs quality parameters recommendations for hard filtering, (iii) had a mapping depth in the range 6–30× and (iv) a *p*-value for excess of heterozygosity greater than 0.05. All SNPs that passed these filtering criteria were used in downstream analyses.

2.3. Targeted Capture Probe Data

Vidalis et al. [22] developed a set 40,018 targeted capture probes that align to exonic and intronic regions in the Norway spruce genome. For all analyses in this paper we use data from the extended probe regions that consist of the 120 bp probes plus 100 bp on either side of the probe, for a total length 320 bp per probe (see Vidalis et al. [22] for more details). SNP data for all extended probe regions were then extracted from the whole-genome re-sequencing data set using BEDTools v2.26.0 [28]. We also extracted data for all scaffolds that were targeted by the capture probes using BEDTools to generate a comparable but more extensive WGS data set to compare with the data from the probe regions. Finally, we supplemented the data for the probe regions extracted from the individuals subjected to re-sequencing with a large data set consisting of 517 individuals previously genotyped using the capture probes [29], to enhance our sample for assessing population structure in Norway spruce.

2.4. Genotyping-by-Sequencing Data

Genotyping-by-sequencing (GBS) was performed using samples of Norway spruce individuals sampled from 45 populations across Västerbotten county in northern Sweden. These samples include 20 of the samples used for whole-genome sequencing (Pab016-Pab035). DNA was extracted from all individuals using the Omega Bio-Tek E-Z 96 plant kit (Omega Bio-Tek, Norcross, GA, USA). Before library preparations 30 samples from each population were pooled in equimolar concentrations with a total of 200 ng (DNA concentrations measured on a Qubit, ThermoFisher Scientific, Waltham, MA, USA) of DNA per sample. To keep the reaction volume down the pools were divided into 8 tubes and dried. Library preparation followed the methods described in Pan et at. [21] with some minor changes outlined below. DNA digestion was made using the restriction enzyme *Pst*I (New England BioLab, Woburn, MA, USA) and during the same time adaptors were ligated to the DNA pools using one of five unique barcodes. Following DNA digestion and ligation, five populations with unique barcodes were pooled producing all together nine independent pools (each containing five populations). All samples were subsequently purified with a QIAquick PCR purification Kit (Qiagen, Hilden, Germany). The DNA was then amplified using a PCR step and purified a second time. Size selection was made with the E-gel Size-Select II pre-cast gel (ThermoFisher Scientific, Waltham, MA, USA) targeting fragments with a size in the range 350–450 bp (including the 125–132 bp barcodes and sequencing adapters). After size selection, DNA was extracted from the gel using QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany). Pair-end sequencing (2 × 150 bp) of all pools were performed on an Illumina HiSeqX by Novogene (Hong Kong) with unique barcodes for all nine pools. Each pool was sequenced individually on one HiSeq X lane per pool, yielding >120 Gbp raw sequencing data per pool.

The raw sequencing data were quality checked using FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and sequencing adaptors were trimmed using Trimmomatic v0.36 [30]. The data was then demultiplexed using the process_radtags routine from Stacks v2.2 [31]. Sequencing reads for all samples were mapped against the *P. abies* v1.0 genome using BWA-MEM. BAM files were

intersected using the multiinter tool from BEDTools v2.26.0 [28] to identify genomic regions common to all samples. SNP data for the genomic regions targeted by the GBS data were then extracted from the whole-genome re-sequencing data set using BEDTools v2.26.0 [28]. As for the probe data, we extracted data for all scaffolds that were targeted by the GBS data using BEDTools to create a comparable data set to compare with the data from the GBS regions.

2.5. Data Analysis

We first estimated relatedness among all 34 individuals using the relatedness2 option in VCFtools [32]. By allowing for the presence of unknown population structure [33], we identified two samples (Pab034 and Pab035) that showed an unusually high degree of relatedness. We therefore excluded one of these samples, Pab034, from all remaining analysis. We then used the 33 remaining individuals, together with the 517 individuals from Baison et al. [29], to assess population structure in the capture probe data to evaluate how much data is needed to capture overall patterns of population structure. Population structure was quantified using a principle component analysis on the relatedness matrix of all samples using the function "prcomp" in R v3.3.3 [34].

In order to determine how well the extended probe regions and the GBS regions capture genetic variation in the WGS data derived from the corresponding scaffold data sets, we estimated a number of summary statistics from the different SNP data sets. For these analyses we only used SNP data from the 19 individuals that were sampled in northern Sweden (Pab016–Pab033 and Pab035). We used ANGSD v0.921 [35] to calculate Tajima's D [36] and nucleotide diversity in the form of pairwise theta [37] for all extended probe regions and for all GBS regions as well as for all probe bearing scaffolds and all GBS scaffolds. We also used VCFtools to calculate F_{ST} values between the four original populations that were sampled in northern Sweden.

3. Results

The extended probe regions cover a total of 12.8 Mb and the corresponding probe-bearing scaffolds cover 567.6 Mb with the probe regions make up approximately 2.3% of the genomic regions encompassed by the probe-bearing scaffolds (Table 2). The GBS regions, on the other hand, cover 1.2 Mb and the scaffolds that encompass the GBS regions have a total size of 171.1 Mb (Table 2), yielding a coverage by the GBS regions of 0.7%. The probe data set cover a larger fraction of coding regions that the GBS data, both for the actual probe regions and the probe bearing scaffolds (Table 2). Similarly, the GBS regions and scaffolds harbor a greater fraction of sequences characterized as repetitive (Table 2). In total, 58,092 bp are common between the extended probe and GBS region data sets and a total of 1965 scaffolds are common between the two scaffold datasets.

Table 2. Summary of probe and GBS data.

	Probes		**GBS**	
	Target Regions	**Scaffolds**	**Target Regions**	**Scaffolds**
Regions	40,018	-	8731	-
Mean size (bp)	320	22,778	140	33,335
Total size (bp)	12,792,815	567,596,728	1,172,909	171,111,561
Unique scaffolds	24,919	-	5133	-
Coding	56.1%	8.5%	20.8%	2.5%
Repeats	1.8%	27.5%	19.0%	38.2%

When analyzing relatedness among the WGS samples, two individuals from the old population at the coast of northern Sweden turned out to be closely related (kinship coefficient $\Phi = 0.526$). To reduce the effect of this on downstream analyses, the sample with the largest amount of missing sequence data (Pab034) was excluded from all subsequent tests.

We used the SNP data for the extended probe regions taken from the WGS dataset to estimate population structure in the 34 WGS samples plus the 517 samples derived from Baison et al. [29]. The results show a clear separation of the samples that mirror the geography of the samples, with clusters representing northern Sweden, Norway, Finland and Central Europe (Figure 1). The sampling is, however, heavily biased towards samples from Scandinavia and Finland and hence best capture variation in the so-called Nordic domain of Norway spruce [38]. This analysis confirms earlier results on the population structure in the re-sequencing samples [24] and earlier results that have shown that stands from southern Sweden contain a large fraction of individuals that have been introduced from central and eastern Europe, including samples that are admixed between central/eastern Europe and southern Sweden (Figure 1, [29,38]). The samples from old growth forests from northern Sweden also separate according to geography (mountains vs. coastal regions, shown as dark green in Figure 1).

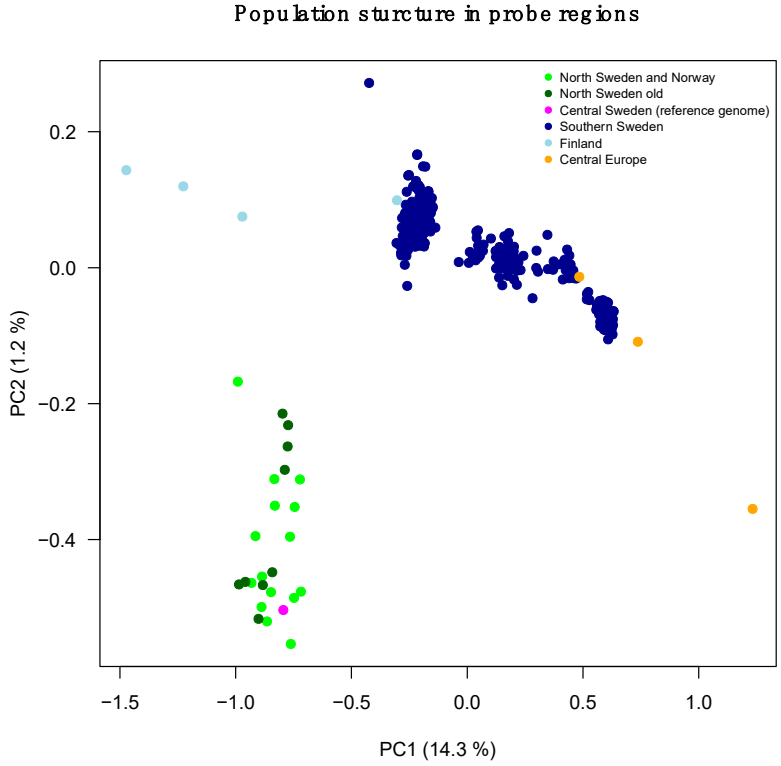

Figure 1. Population structure in probe regions.

The capture probes were initially selected to have intermediate levels of genetic variation [22] and this will hence result in a certain degree of ascertainment bias in the polymorphism data generated from the extended probe regions. Tajima's D does not appear to differ in any systematic way between the extended probe regions and the data generated from the entire probe-bearing scaffold set (Figure 2A, Table 3). For nucleotide diversity (pairwise theta), scaffolds appear to be slightly more variable on average than the extended probe regions, but overall the extended probe regions accurately capture both the amount and frequency spectrum of genome-wide nucleotide polymorphism.

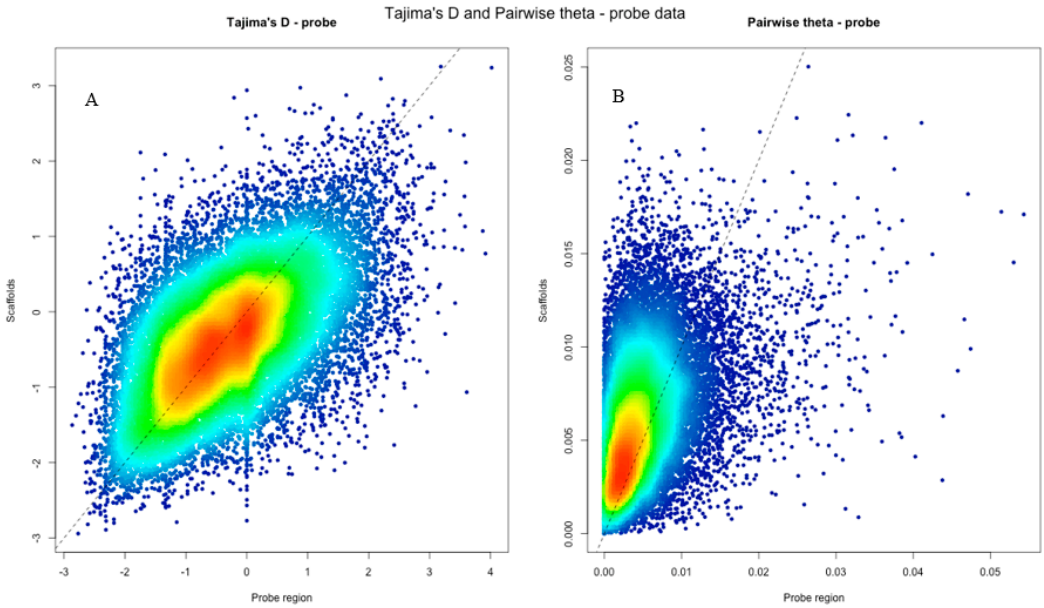

Figure 2. (**A**) Tajima's D and (**B**) pairwise theta (nucleotide diversity) in the probe data. Red indicates the highest number of data points and dark blue the lowest.

Table 3. Median and standard deviation (in parenthesis) for nucleotide diversity and Tajima's D for the different data sets.

Subset	Pairwise Theta	Tajimas'D
Probe scaffolds	0.0056 (0.0032)	−0.376 (0.84)
Probe regions	0.0040 (0.0044)	−0.36 (1.017)
GBS scaffolds	0.0063 (0.0031)	−0.393 (0.792)
GBS regions	0.0014 (0.0034)	−0.311 (1.003)

The GBS data display a similar pattern, only with slightly fewer extreme values in Tajima's (Figure 3A, Table 3). There are fewer data points in Figure 3 than in Figure 2, simply because the GBS data contain a substantially smaller number of unique regions compared to the extended probe regions (Table 2). While GBS regions have lower diversity values than their corresponding scaffolds (Figure 3B), the site frequency spectrum summarized by Tajima's D, does not show any appreciable differences overall between regions and scaffolds (Figure 3A).

Genetic differentiation between populations also show little evidence for any systematic differences between targeted regions and scaffolds in both probe and GBS data sets (Figures 4 and 5, [39]). Short probe or GBS regions thus appear to accurately capture genome-wide variation in genetic differentiation. Population 2.1 is an old costal population and 2.3 a young, recently planted costal population. Similarly, population 12.1 is an old population sampled from the mountain region in Västerbotten county and 12.3 is a newly planted population also located in the mountain region. F_{ST} values between the different pairs of populations from northern Sweden all show very weak population differentiation independent on the data source and populations that are compared (median F_{ST} for probe data is −0.019–−0.0018, median F_{ST} for GBS data is −0.018–0.007). However, even if the average genetic differentiation among populations is low, there is abundant variation in genetic differentiation across the Norway spruce genome, as many genomic regions in both the probe and GBS data that show substantial genetic differentiation (F_{ST} values > 0.2).

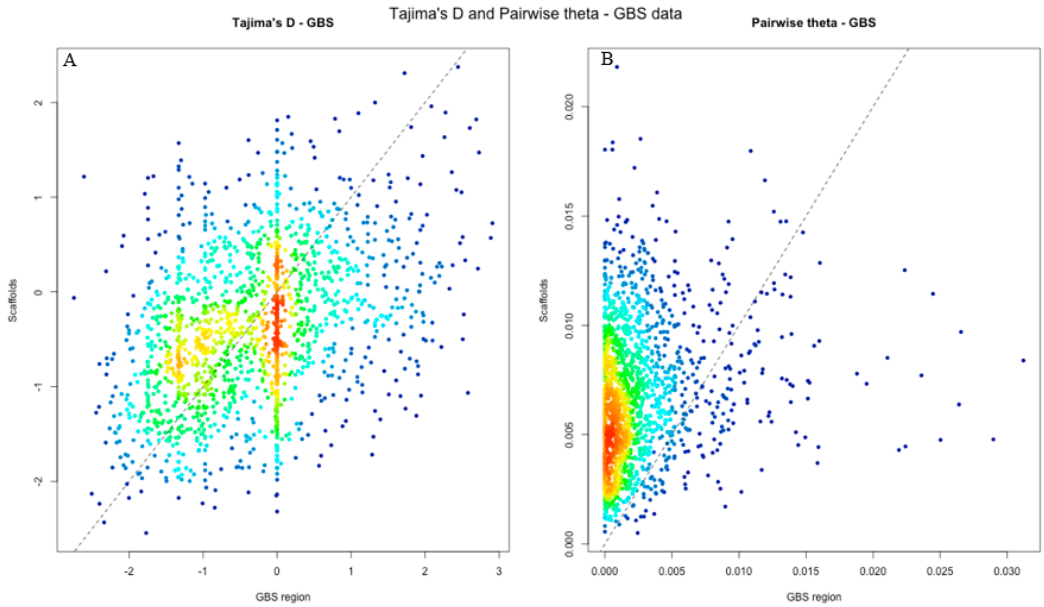

Figure 3. (**A**) Tajima's D and (**B**) pairwise theta (nucleotide diversity) in the GBS data. Red indicates the highest number of data points and dark blue the lowest.

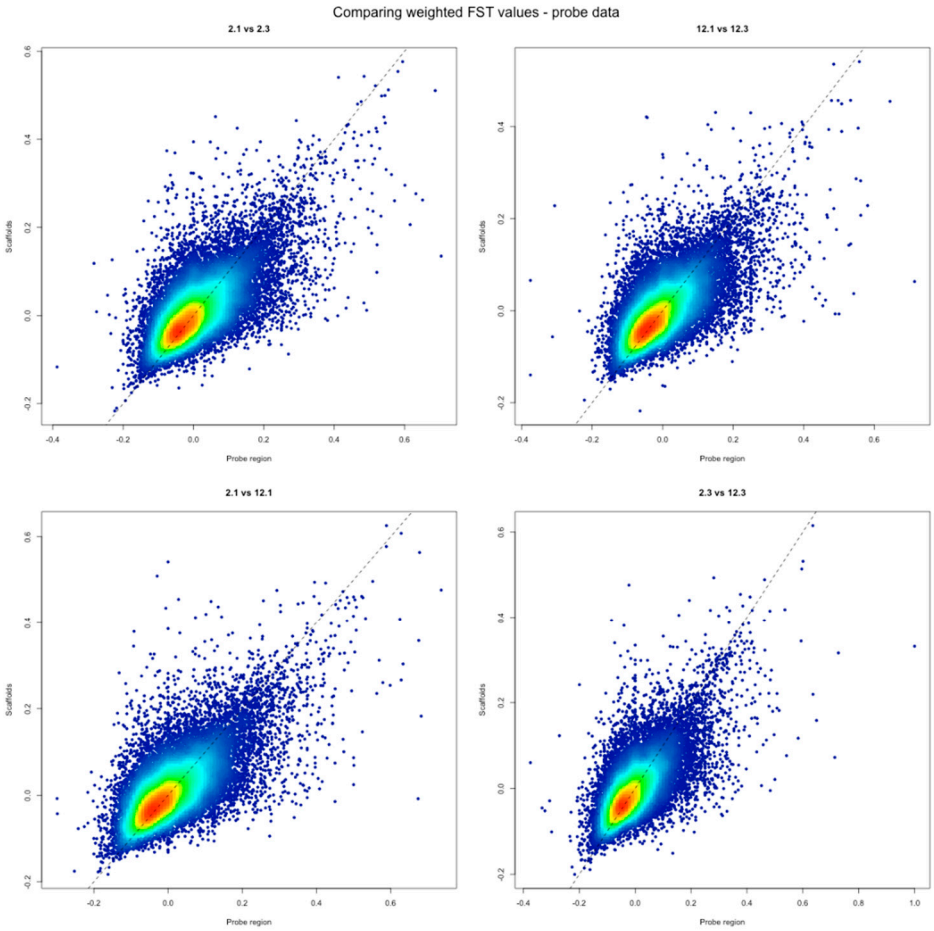

Figure 4. Comparing Fst values between populations in the probe data. Population 2.1 is an old costal population and 2.3 a young, recently planted costal population. 12.1 is an old population from the mountain region in Västerbotten county and 12.3 is a newly planted population in the mountain region. Red indicates the highest number of data points and dark blue the lowest.

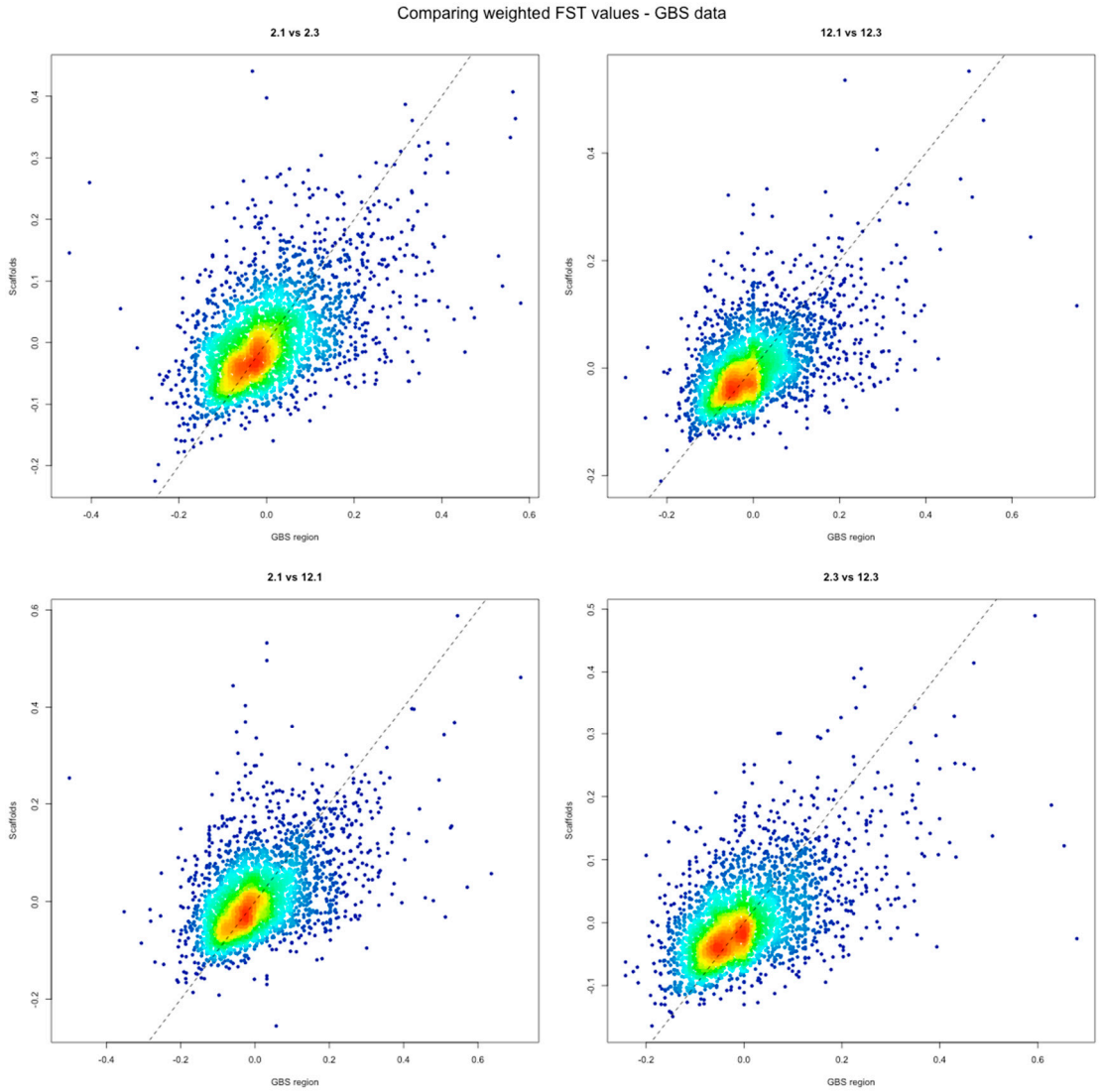

Figure 5. Comparing Fst values between populations in the GBS data. Population 2.1 is an old costal population and 2.3 a young, recently planted costal population. 12.1 is an old population from the mountain region in Västerbotten county and 12.3 is a newly planted population in the mountain region. Red indicates the highest number of data points and dark blue the lowest.

4. Discussion

In this study we found no large differences when using either WGS, capture probe data or GBS data for assessing genetic diversity in Norway spruce. The main difference between the two complexity reduction methods is what prior knowledge that is needed to develop and use a given method. Sequence capture probes are designed to target known regions in a genome and hence require access to either a draft genome or at least extensive sequence information, such as RNAseq or other targeted sequencing, from an organism in order to design them before use. GBS, on the other hand, can be applied to any organism without any prior genome information. Complexity reduction with GBS is only based on selecting a specific fragment size from a gel and is thus more 'random' with respect to the genomic regions it targets and should therefore, at least in theory, provide a more accurate representation of the genome in an organism. In our data this is most apparent from higher repetitive content contained in the different sequence data sets related to the GBS method compared to the sequence capture probe method (Table 2).

Comparing estimates of genetic diversity for the different datasets show that the scaffold data sets had less extreme values, both high and low, compared to the reduced representation regions (capture probes or GBS, Figures 2A and 3A), a simple consequence of the greater size of regions covered in the scaffold data. This makes estimates of diversity in these data sets less prone to stochastic variation. For the two complexity reduction data sets, measurements of diversity GBS seems to be slightly less effective than probe data and this is likely a consequence of the method itself. Using a restriction enzyme to perform complexity reductions does not allow for precise control over where the DNA gets cut. For example, if a target cut site region has high nucleotide diversity it is possible that there is also polymorphism at sites in the base pairs making up the cut site, effectively reducing the number of cuts and also generating differences among individuals in what fragments gets included in the final GBS data. Since GBS relies on selecting a predetermined fragment size, fragments that lack a specific cut site will generate fragment sizes outside of the selection range for some individuals and this ultimately leads to missing data between different individuals and also to a relatively small overlap across individuals if no missing data is accepted. Depending on how much missing data that is allowed, the overlap of regions across individuals will vary since missing data will create regions where data is not present in every individual. This is simply a consequence of the GBS method and partly explain why the size of genomic regions covered in the GBS data is relatively small, only 1.2 Mb, when we require no missing data across individuals. If one allows for 10% or 20% missing data in the GBS results, the size of the regions covered increase to 1.4 Mb and 1.6 Mb, respectively. Repetitive regions are also problematic for downstream genotyping [25], regardless of method and the greater fraction of repetitive regions in the GBS could hence further skew the results from these regions. The overall summary of the data (Table 2) show that the capture probes generates more data both in terms of total size and number of unique scaffolds and also, as expected, contain a greater fraction coding regions.

Using only capture probe data we were also able to reconstruct the population structure of Norway spruce trees from Sweden and the rest of Europe with high accuracy and thereby also confirming earlier results from Wang et al. [24]. For assessing population structure and genetic differentiation a relatively modest amount of data is required to accurately capture the overall patterns in the data. In the analyses of genetic differentiation among populations, both the capture probe and GBS data yield the same general results, although the GBS is slightly more variable. This is again likely caused by the lower amount of data available in the GBS data set. However, even though the GBS data has more extreme values the median F_{ST} values are similar for all data sets. It is also worth noticing that even if median values for F_{ST} are very low and close to zero for all data sets, there are still regions showing strong genetic differentiation with F_{ST} values above 0.2 (Figures 4 and 5).

5. Conclusions

Our results show that both capture probes and GBS are performing reasonably well for assessing genetic diversity and population structure in Norway spruce. Targeted capture probes are slightly more effective and moderately variable than GBS, by virtue of targeting regions known *a priori* to be largely unique in the Norway spruce genome while the GBS data contains a substantially greater fraction of repetitive regions. This is not surprising since GBS selects regions anonymously and should hence more accurately reflect the high repetitive content present in Norway spruce [14]. However, GBS methods can be tailored to reduce the fraction of repetitive regions targeted, for example by using methylation sensitive restriction enzymes that allows for enrichment of unique genomic regions, such as coding regions [40]. Nevertheless, both capture probes and GBS yield the same results on average. These methods are hence useful when genotyping large numbers of samples and they greatly reduce the cost involved with genotyping a species with such a complex genome as Norway spruce. Our research is focused on large geographic areas with differences in both biotic and abiotic factors. For us to be able to show possible differences caused by forestry practices we have to be able to genotype a large number of samples, both planted and from natural stands from every area. With a cost-efficient option like GBS we now have the tools to conduct genotyping on multiple stands of Norway spruce.

Author Contributions: Conceptualization, P.K.I; methodology and formal analysis, H.E., C.B. and P.K.I., data curation, H.E. and C.B.; writing—original draft preparation, H.E.; writing—review and editing, C.B. and P.K.I.; supervision, C.B. and P.K.I. All authors have read and agreed to the published version of the manuscript.

Funding: This study has been supported by grants from the Knut and Alice Wallenberg Foundation (KAW) and the Swedish Foundation for Strategic Research (SSF), grant number RBP14-0040.

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

References

1. Källman, T. Adaptive Evolution and Demographic History of Norway Spruce (*Picea abies*). Ph.D. Thesis, Uppsala University, Uppsala, Sweden, 2009.
2. Farjón, A. *Pinaceae: Drawings and Descriptions of the Genera Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix and Picea*; Koeltz Scientific Books: Koönigstein, Germany, 1990.
3. Burczyk, J.; Lewandowski, A.; Chalupka, W. Local pollen dispersal and distant gene flow in Norway spruce (*Picea abies* [L.] Karst.). *For. Ecol. Manag.* **2004**, *197*, 39–48. [CrossRef]
4. Lindgren, D.; Karlsson, B.; Andersson, B.; Prescher, F. Swedish seed orchards for Scots pine and Norway spruce. In Proceedings of the a Seed Orchard Conference, Umeå, Sweden, 26–28 September 2007; pp. 26–28. Available online: http://daglindgren.upsc.se/Umea07/Umea07.htm (accessed on 6 November 2020).
5. Lindgren, D.; Prescher, F. Optimal clone number for seed orchards with tested clones. *Silvae Genet.* **2005**, *54*, 80–92. [CrossRef]
6. Adams, W.T.; Burczyk, J. Magnitude and implications of gene flow in gene conservation reserves. In *forest Conservation Genetics: Principles and Practice*; Young, A., Boshier, D., Boyle, T., Eds.; CSIRO Publishing: Canberra, Australia, 2000; pp. 215–244.
7. Pakkanen, A.; Nikkanen, T.; Pulkkinen, P. Annual variation in pollen contamination and outcrossing in a *Picea abies* seed orchard. *Scand. J. For. Res.* **2000**, *15*, 399–404. [CrossRef]
8. Paule, L.; Lindgren, D.; Yazdani, R. Allozyme frequencies, outcrossing rate and pollen contamination in *Picea abies* seed orchards. *Scand. J. For. Res.* **1993**, *8*, 8–17. [CrossRef]
9. Rosvall, O.; Almqvist, C.; Lindgren, D.; Hallander, J.; Berlin, M. Updates from Research on Selection and Mating Strategies. Review of the Swedish tree Breeding Programme. Skogforsk. 2011. Available online: https://www.skogforsk.se/kunskap/kunskapsbanken/2011/Utvardering-av-Skogforsks-foradlingsstrategi/ (accessed on 6 November 2020).
10. Scotti, I.; Gugerli, F.; Pastorelli, R.; Sebastiani, F.; Vendramin, G.G. Maternally and paternally inherited molecular markers elucidate population patterns and inferred dispersal processes on a small scale within a subalpine stand of Norway spruce (*Picea abies* [L.] Karst.). *For. Ecol. Manag.* **2008**, *255*, 3806–3812. [CrossRef]
11. Mehra, P.N.; Khoshoo, T.N. Cytology of conifers. I. *J. Genet.* **1956**, *54*, 165–180. [CrossRef]
12. Morse, A.M.; Peterson, D.G.; Islam-Faridi, M.N.; Smith, K.E.; Magbanua, Z.; Garcia, S.A.; Kubisiak, T.L.; Amerson, H.V.; Carlson, J.E.; Nelson, C.D.; et al. Evolution of genome size and complexity in *Pinus*. *PLoS ONE* **2009**, *4*, e4332. [CrossRef]
13. Neale, D.B.; Wheeler, N.C. The Conifers. In *The Conifers: Genomes, Variation and Evolution*; Springer: Cham, Switzerland, 2019; pp. 1–21.
14. Nystedt, B.; Street, N.R.; Wetterbom, A.; Zuccolo, A.; Lin, Y.C.; Scofield, D.G.; Vezzi, F.; Delhomme, N.; Giacomello, S.; Alexeyenko, A.; et al. The Norway spruce genome sequence and conifer genome evolution. *Nature* **2013**, *497*, 579–584. [CrossRef]
15. Heuertz, M.; De Paoli, E.; Källman, T.; Larsson, H.; Jurman, I.; Morgante, M.; Lascoux, M.; Gyllenstrand, N. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* **2006**, *174*, 2095–2105. [CrossRef]
16. Prunier, J.; Laroche, J.; Beaulieu, J.; Bousquet, J. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Mol. Ecol.* **2011**, *20*, 1702–1716. [CrossRef]
17. Acheré, V.; Favre, J.M.; Besnard, G.; Jeandroz, S. Genomic organization of molecular differentiation in Norway spruce (*Picea abies*). *Mol. Ecol.* **2005**, *14*, 3191–3201. [CrossRef] [PubMed]

18. Gapare, W.J.; Aitken, S.N.; Ritland, C.E. Genetic diversity of core and peripheral Sitka spruce (*Picea sitchensis* (Bong.) Carr) populations: Implications for conservation of widespread species. *Biol. Conserv.* **2005**, *123*, 113–123. [CrossRef]
19. Chen, C.; Mitchell, S.E.; Elshire, R.J.; Buckler, E.S.; El-Kassaby, Y.A. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genet. Genomes* **2013**, *9*, 1537–1544. [CrossRef]
20. Karam, M.-J.; Lefèvre, F.; Dagher-Kharrat, M.B.; Pinosio, S.; Vendramin, G.G. Genomic exploration and molecular marker development in a large and complex conifer genome using RADseq and mRNAseq. *Mol. Ecol. Resour.* **2015**, *15*, 601–612. [CrossRef] [PubMed]
21. Pan, J.; Wang, B.; Pei, Z.Y.; Zhao, W.; Gao, J.; Mao, J.F.; Wang, X.R. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. *Mol. Ecol. Resour.* **2015**, *15*, 711–722. [CrossRef] [PubMed]
22. Vidalis, A.; Scofield, D.G.; Neves, L.G.; Bernhardsson, C.; García-Gil, M.R.; Ingvarsson, P.K. Design and evaluation of a large sequence-capture probe set and associated SNPs for diploid and haploid samples of Norway spruce (*Picea abies*). *bioRxiv* **2018**, 291716. [CrossRef]
23. De La Torre, A.R.; Puiu, D.; Crepeau, M.W.; Stevens, K.; Salzberg, S.L.; Langley, C.H.; Neale, D.B. Genomic architecture of complex traits in loblolly pine. *New Phytol.* **2019**, *221*, 1789–1801. [CrossRef]
24. Wang, X.; Bernhardsson, C.; Ingvarsson, P.K. Demography and natural selection have shaped genetic variation in the widely distributed conifer Norway spruce (*Picea abies*). *Genome Biol. Evol.* **2020**, *12*, 3803–3817. [CrossRef]
25. Bernhardsson, C.; Wang, X.; Eklöf, H.; Ingvarsson, P.K. Variant calling using whole genome resequencing and sequence capture for population and evolutionary genomic inferences in Norway Spruce (*Picea abies*). In *The Spruce Genome. Compendium of Plant Genomes*; Porth, I., De la Torre, A., Eds.; Springer: Cham, Switzerland, 2020; pp. 9–36. [CrossRef]
26. 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]
27. 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]
28. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinform.* **2014**, *47*, 11–12. [CrossRef] [PubMed]
29. Baison, J.; Vidalis, A.; Zhou, L.; Chen, Z.-Q.; Li, Z.; Sillanpää, M.J.; Bernhardsson, C.; Scofield, D.; Forsberg, N.; Grahn, T.; et al. Genome-wide association study (GWAS) identified novel candidate loci affecting wood formation in Norway spruce. *Plant J.* **2019**, *100*, 83–100. [CrossRef] [PubMed]
30. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
31. Catchen, J.M.; Amores, A.; Hohenlohe, P.; Cresko, W.; Postlethwait, J.H. Stacks: Building and genotyping loci de novo from short-read sequences. *G3* **2011**, *1*, 171–182. [CrossRef]
32. 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]
33. Manichaikul, A.; Mychaleckyj, J.C.; Rich, S.S.; Daly, K.; Sale, M.; Chen, W.M. Robust relationship inference in genome-wide association studies. *Bioinformatics* **2010**, *26*, 2867–2873. [CrossRef]
34. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017. Available online: https://www.R-project.org/ (accessed on 6 November 2020).
35. Korneliussen, T.S.; Albrechtsen, A.; Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinform.* **2014**, *15*, 356. [CrossRef]
36. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **1989**, *123*, 585–595.
37. Korneliussen, T.S.; Moltke, I.; Albrechtsen, A.; Nielsen, R. Calculation of Tajima's *D* and other neutrality test statistics from low hh next-generation sequencing data. *BMC Bioinform.* **2013**, *14*, 289. [CrossRef]

38. Chen, J.; Li, L.; Milesi, P.; Jansson, G.; Berlin, M.; Karlsson, B.; Aleksic, J.; Vendramin, G.G.; Lascoux, M. Genomic data provides new insights on the demographic history and the extent of recent material transfers in Norway spruce. *Evol. Appl.* **2019**, *12*, 1539–1551. [CrossRef]
39. Shimono, A.; Wang, X.R.; Torimaru, T.; Lindgren, D.; Karlsson, B. Spatial variation in local pollen flow and mating success in a *Picea abies* clone archive and their implications for a novel "breeding without breeding" strategy. *Tree Genet. Genomes* **2011**, *7*, 499–509. [CrossRef]
40. Fellers, J.P. Genome filtering using methylation-sensitive restriction enzymes with six base pair recognition sites. *Plant Genome* **2008**, *1*, 146–152. [CrossRef]