Ref. Ares(2021)5122036 - 16/08/2021

B4EST

Adaptive BREEDING for productive, sustainable and resilient FORESTs under climate change

Deliverable D3.1

Multispecies SNP array as a common genotyping tool for breeding

Planned delivery date (as in DoA): M20

Actual submission date: 06/08/21 (M40)

Workpackage: WP3

Workpackage leader: INRAE

Deliverable leader: CNR

Version: 2.0

Project co-funded by the European Commission within the Horizon 2020 Programme (2014 - 2020)	
Dissemination Level	
PU Public	**PU**
CI Classified, as referred to Commission Decision 2001/844/EC	
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation action: GA no. 773383

Start date of the project: May 1st, 2018

TABLE OF CONTENTS

1 Summary 4
2 Introduction 4
3 A 420K Screening array to design a 50K array for 4TREE array 5
3.1 SNP calling and filtering 5
3.1.1 Pinus pinea 5
3.1.2 Pinus pinaster 6
3.1.3 Populus sp 6
3.1.4 Fraxinus sp 6
3.2 Axiom SNP scoring results 7
3.3 Screening array development 7
3.3.1 Pinus pinea 7
3.3.2 Pinus pinaster 8
3.3.3 Populus sp 8
3.3.4 Fraxinus sp 9
3.4 Genotype panel for validation of the screening array 9
4 Multiple species array 4TREE 10
4.1 Quality controls of the screening array 10
4.2 Selection of informative and robust SNP in Pinus pinea 11
4.3 Selection of informative and robust SNP in Pinus pinaster 12
4.4 Selection of highly informative SNP for Populus deltoides and P. nigra 12
4.5 Selection of informative and robust SNP in Fraxinus sp 14
4.6 Summary for 4TREE array 15
5 Selection of informative and robust SNPs in *Picea abies* 15
5.1 Construction of the 450K pilot screening array 15
5.2 Selection of the 50K SNP array from the pilot screening array 16
5.3 Evaluation and validation of the 50K genotyping array 16
6 Development of a species-specific genotyping array for *Pinus sylvestris* from screening array to final PiSy50k array 17
7 Conclusion 18
8 Partners involved in the work 20

9 Annexes 21

10 Bibliography 31

1 Summary

The main challenges facing WP3.1 was how to choose an appropriate SNP (Single Nucleotide Polymorphism) genotyping method to create useful, affordable, high-throughput, robust (cross platform utilization), efficient genotype calling software and largely reusable genotyping tool. Three Axiom arrays are currently under development, the first one is a multispecies array (4TREE) created for *Populus* sp, *Fraxinus* sp, *Pinus pinaster* and *Pinus pinea*. It is based on *de novo* sequencing resources and existing genomic resources of loci of interest for breeders. With 4TREE, about 12.5K SNPs of good quality for each species will be available. These SNPs are selected from 420K SNPs evaluated on a set of 120 individuals per species. 4TREE is primarily aimed at breeders by providing low cost genotyping methods for economically important species in Europe, fostering collaboration between breeding programs across Europe. The second (Pcab) array is a 50K SNP array designed for *Picea abies* thanks to a 450K pilot array based on 29 whole genome re-sequenced Norway spruce samples. The third one (PiSy50K) designed for *Pinus sylvestris* was developed using multisource SNP. One of first uses of these three arrays will be genome wide association studies and genomic predictions in the frame of B4EST project. These genotyping tools which will be made commercially available worldwide will be also extremely useful for monitoring and characterizing novel sources of variation in genetic conservation or breeding programs.

2 Introduction

The SNP genotyping activity in plant breeding is nowadays an essential tool for optimizing selection schemes. This is even more important in forest species that are non-domesticated perennial species with late expression of phenotypic traits. Since the early 2000s, with high-throughput genomics development, many possibilities have emerged to create genotyping tools. Two main categories of SNP genotyping platform exist: SNP array and Genotyping by Sequencing (GBS). In this last category, there are several possible approaches: whole genome sequencing, reduced representation sequencing or targeted sequencing. Each platform and approach has advantages and disadvantages that are detailed in several review papers (Scheben et al. 2017; Rasheed et al. 2017; Torkamaneh et al. 2018). In the frame of B4EST, we have chosen to develop a SNP array, taking into account in one hand advantages of the tool (robust, transferable, user-friendly genotype calling system), and on the other hand technical and scientific expertise of the B4EST partners. We started negotiation with the two array providers: Illumina (http://www.illumina.com) and ThermoFisher Scientific (http:// www.thermofisher.com). Both technologies have proven successful for genetic diversity, genetic and QTL mapping (Quantitative Trait Loci), GWAS (Genome Wide Association Studies) and genomic selection in several forest tree species (Annex 1).

Because the genotyping cost decreases with the number of samples and the number of SNPs analyzed, all the partners have decided during the kick off meeting (Orléans, June 2018) to join the international consortium named CSC (Conifer SNP Consortium led by North Carolina State University (F. Isik)). All together, we increased the demand of forest tree genotyping and beneficiated of a substantial reduced cost. This consortium started discussions with the two suppliers of SNP array. The proposal made by ThermoFisher was by far the most interesting for the entire consortium including the B4EST partners. The tool proposed by Thermofisher scientific (Axiom array) offers an

advantage that was not offered by Illumina (Infinium array) consisting in a preliminary round of genotyping (pre-screening) to optimize the design of the final genotyping tool. Thus, the first round allows us to identify the most robust and informative SNPs, by testing a subset of trees, to include in a final array. Therefore, technicities and rates made the ThermoFisher Scientific option the preferred one and the one selected by the B4EST consortium. Finally, taking into account the scientific and socio-economic context of the improvement programs for each species, 3 ThermoFisher tools are being developed. A chip specific to *Picea abies* (coordination H. Wu from SLU, SE) and a multispecies array including SNPs for *Pinus pinea*, *Pinus pinaster*, *Populus sp* and *Fraxinus* sp (named 4TREE, P. Faivre Rampant from INRA, FR) are developed within the CSC. The last one is developed outside the CSC for *Pinus sylvestris* (coordination T. Pyhäjärvi from OULU, FI and S. Cavers from NERC, UK), but still using the same technique and provider.

This deliverable is reporting both the construction of the first array named “screening array” of 420K SNPs to help the design of the 4TREE 50Karray and presentation of the three final arrays.

3 A 420K Screening array to design a 50K array for 4TREE array

3.1 SNP calling and filtering

3.1.1 Pinus pinea

Pinus pinea has been cultivated for thousands of years and is economically important for its nuts but its breeding is limited (Mutke et al. 2012). In a genomic context, *Pinus pinea* is among the so-called orphan species as no reference genome is available and few genomic studies have been performed (Alonso et al. 2007; Vendramin et al. 2010). Genomic resources were created inside B4EST to identify SNPs. Five individuals originated from Spain, Greece, Lebanon, Turkey and Morocco were selected to identify SNPs though RNA sequencing. RNA extraction and library preparation were performed according to standard protocols and carried out at IGATech (UDINE I). Sequencing was done on Hiseq2500 (Illumina, San Diego, CA) with a 125 bp paired-end mode. The number of reads was variable between samples and ranged from 73 million up to 99 million.

After trimming and filtering, high quality reads were aligned to the *Pinus taeda* v2.0 reference genome (Zimin et al. 2017) using the RNA-seq aligner STAR v2.6 (Dobin et al. 2013) by setting outFilterMultimapNmax parameter to 5 and outFilterMismatchNmax to 6. After alignment, only reads with a mapping quality of at least 10 were selected using samtools (Li 2011) and duplicated sequences were removed using the picard tools MarkDuplicates utility (https://broadinstitute.github.io/picard/). Aligned reads were split into exon segments and sequences overhanging into the intronic regions were hard-clipped using the tool SplitNCigarReads of GATK v4.0.3.0 (DePristo et al. 2011). SNP calling was performed using the tool HaplotypeCaller of GATK v4.0.3.0 (DePristo et al. 2011) with default parameters. SNPs were selected using GATK SelectVariants and filtered using GATK VariantFiltration by applying the following filter expression: QD < 2.0 || FS > 30.0. For each sample, SNPs were additionally filtered by requiring a minimum coverage of 8 and a minimum percentage of reads supporting the alternative allele 20%. Finally, interspecific SNPs (i.e. *P. pinea* vs. *P. taeda* polymorphisms) were discarded and only positions informative in at least two samples were retained. A total of 128 279 SNPs with 35 bp up and down stream flanking sequences were obtained and delivered to ThermoFisher for the scoring process.

3.1.2 *Pinus pinaster*

Most of SNPs come from genomic resources created last year within the H2020 GenTree project. The SNPs were discovered from exome capture experiment carried out from a diverse panel of 163 individuals with the same procedure described for *P. pinea* (paragraph 3.1.1). Moreover, 5 165 high quality SNPs from the 9K Illumina Infinium array generated by Plomion et al. (2016) were added. In total, 191 785 SNPs were sent for scoring.

3.1.3 *Populus sp*

Two main species are involved in current *Populus* breeding programs in Italy and France: the European species *Populus nigra* and the American one *Populus deltoides*. Different categories of SNPs were searched: substitutions between *P. nigra* and *P. deltoides* to check hybrids, SNPs specific to each species and shared SNPs between species.

The SNPs data for *P. nigra* were discovered through various European and national projects and from literature (Evoltree, NovelTree, Sybiopop, EpiTree, Faivre-Rampant et al. 2016; Rogier et al. 2018; Pégard et al. 2019). In the frame of the national projects BreedToLast and EpiTree, whole genome sequencing was carried out to discover SNPs. Respectively 13 448,600 and 12 432 187 SNPs were from these projects. Moreover, 878 827 additional SNPs were identified from RNA sequencing on 241 individuals from the project Sybiopop.

P. deltoides SNP discovery were performed inside B4EST. Available genomic resources such as whole genome sequencing from 5 individuals (Accession number: SRS 1320691, SRS 1328499, unpublished data) and exome capture sequencing data performed on 579 were used to detect SNPs specific to *P. deltoides* (Fahrenkrog et al. 2017). The method of SNP calling has been described in a previous study of Rogier et al, 2018. Combining the different data sets, 366 807 SNPs were detected.

The SNPs were selected to have a balance between: biallelic SNPs with alleles specific to each species, biallelic SNPs with alleles shared between both species, SNPs well distributed in the genome, a 60/40 ratio of genic/intergenic DNA. Additionally, high quality SNPs from the existing array (8 500 SNPs, Faivre-Rampant et al. 2016) and candidate SNPs obtained through a genome wide association study (V. Segura, unpublished data) were also preselected. A total of 2 213 994 SNPs were sent to ThermoFisher for scoring.

3.1.4 *Fraxinus* sp

The SNPs filtered for inclusion in the SNP array coming mainly from published data (Sollars et al. 2017; Stocks et al. 2019; Kelly et al. 2019). Two thousand and five hundred SNPs associated to ash dieback disease; 64 associated to emerald ash borer susceptibility and 26 SNPs located in candidate regions linked to incompatibility system of the *oleaceae* family. After filtering out triallelic SNPs and multiple hits on the BAT-G.0.5 ash genome used as reference, 2 474 SNPs remained. Other SNPs (324 590) were detected using associative transcriptomic data. In this set, CG and AT variants, triallelic loci and multiple hits on the BAT-G.0.5 genome were discarded. At the end, a total of 327 064 SNPs were sent for scoring to ThermoFisher.

3.2 Axiom SNP scoring results

ThermoFisher Bioinformatics service scored the capability to design probes in the SNP flanking sequences on both the forward strand and the reverse strand. For each strand, a recommendation is given, which is relative to the ability of the SNP/probes to be tiled. The recommendation levels are as follows:

- A SNP is considered as "**recommended**" if the pconvert (probability of conversion) > 0.6, there are no interfering polymorphisms within 24 bases and if the 24 bases flanking sequences are not shared with another SNP.

- A SNP is "**not recommended**" if the SNP and its flanking sequences are duplicated in the list (duplicate_count), or if the 24 bases in flanking sequences are shared with another SNP (probe_count), or if the pconvert<0.4, or if there are interfering polymorphisms in the flanking sequences (wobble_count>0), or if there are homology problems (homology_count).

- A SNP is "**not possible**" if a probe cannot be built. In our case, all the SNPs with a strand "not possible" could have a probe made on the other strand.

- All the other markers are "**neutral**"

The scoring by ThermoFisher on the SNPs of the four species is given in the table 1.

Table 1. Scoring results of submitted SNP

	Populus	***Fraxinus***	***Pinus pinea***	***Pinus pinaster***
SNPs submitted	2 213 994	327 064	128 279	191 785
Recommended	1 738 409	279 337	3 495	6 562
Neutral	196 596	26 059	3 212	7 898
Not recommended	278 989	21 668	121 572	177 325

The lower rate of recommended probes obtained for the pine species is typical for a large genome (>15 Gb) (CSC communication). Nevertheless, it can be noticed that around 3/5 of the successful SNPs from the Illumina Infinium array were classified as "not recommended". Taking into account the good rate of transferability between the two technologies Infinium and Axiom (D. Grattapaglia, personal communication), successful SNPs among those "not recommended" can be expected.

3.3 Screening array development

3.3.1 Pinus pinea

Due to the number of not recommended (94,8%), numerous SNPs have been eliminated during the selection. Consequently, only 36 784 markers were kept for the screening array (the remaining

slots were distributed among the other species). The SNPs from *Pinus pinea* were classified into four priority levels:

- Priority 1 comprises all the recommended, neutral SNPs and isolated markers with flanking sequences having a good sequence homology (homology_count < 30). There were 11 861 SNPs priority 1.

- The rest of the isolated SNPs and markers with a probability of conversion superior to 0 (that have a chance to work) were included in the priority 2 level. A total of 11 957 SNPs were classified as priority 2.

- Among the rest of the markers, those that had unique probes, no interfering polymorphisms in the flanking sequences and good homology (wobble_count=0, homology_count<30 and probe_count="empty") were classified in priority 3 level. This concerns 7 395 SNPs.

- The markers with unique probes, no interfering polymorphisms in the flanking sequences (wobble_count=0 and probe_count="empty") but failed the homology criteria for priority 3 were put in priority 4. There were 5 574 SNPs in priority 4.

Almost all of these markers (36 784) were kept in the screening array design, including 5 871 SNPs A/T or C/G.

3.3.2 *Pinus pinaster*

Among the 191 785 SNPs, 137 321 SNPs were kept for the screening array. Five priority levels have been defined for *Pinus pinaster*.

- Priority 1 markers include the 5 163 SNPs from the existing Illumina Infinium array.

- All the other markers were distributed inside the four remaining priorities. The markers that were recommended (3,4% of the markers) or neutral (4,1%) by the ThermoFisher scoring were in priority 2. This concerns 12 572 SNPs.

- The 31 106 SNPS with a probability of conversion superior to 0,3 were in priority 3.

- The priority 4 consisted of the 38 608 SNPs with no interfering polymorphisms and had unique probes (wobble_count=0 and probe_count="empty").

- Finally, the 52 662 SNPs priority 5 SNPs were markers with only a few mismatch in the probe (homology_count<=20)

Among the selected 137 321 SNPs, 17 857 were A/T or C/G. Five thousand SNPs from the existing Infinium SNP were tiled on forward and backward strand to increase their chances of success.

3.3.3 *Populus* sp

Most of the SNP were recommended (78,5%), i.e. 1 738 409 markers, while 196 596 were neutral and 278 989 were not recommended.

A selection of 120 504 markers was made for the screening array. The priority was to include intraspecific SNPs to one or both species with a MAF>5%, but also biallelic interspecific SNPs

(substitutions). The integration of previous data such as SNPs from the existing Infinium array or candidate SNPs from GWAS was also a priority. Another main criterion was a homogeneous coverage over the genome and a 60/40 genic/intergenic regions ratio.

Moreover to fill the array, ThermoFisher proposed additional SNP to reach the number of 137 545 SNPs. About 4 000 of these markers used two rooms in the array as they are A/T or C/G variants.

3.3.4 Fraxinus sp

For *Fraxinus* sp, 327 064 SNPs were submitted, most of them were recommended (85,4%), 26 059 were neutral and 21 668 were not recommended

A selection of 126 000 SNPs was made with three priority levels.

- Priority 1 markers were the candidates associated with ash dieback disease loci emerald ash borer susceptibility and incompatibility (Stocks et al. 2019), up to 8 SNPs per gene or transcript. Priority 1 level includes 2 474 SNPs.

- Priority 2 contains up to 8 SNPs per gene for the SNPs from associative transcriptomics with a minor allele frequency higher than 5%. There were 62 713 SNPs in this level.

- Priority 3 SNPs were taken from the recommended markers in the contigs not represented in priorities 1 and 2. They were grouped into two subcategories. The category 3a contains one SNP from each of the 3 651 contigs not in 2 and three SNPs from each of the 29 534 contings not in 1. Only one SNP was selected when the contig had less than 3 recommended SNPs; these SNPs were included in priority 3b. It concerns 4 033 SNPs.

All these SNPs (138 038) covered 48 742 contigs, covering a total of 753,8 Mb of the *Fraxinus* genome, 86,9% of the total genome assembly (Sollars et al. 2017).

At the end the screening array contains 449 688 SNP (Table 2).

3.4 Genotype panel for validation of the screening array

A number of 546 trees (120 *Pinus pinea*, 120 *Pinus pinaster* 128 Populus sp, 176 *Fraxinus* sp, Annex 2) were selected to represent the diversity used in European breeding programs. Total DNAs were extracted and sent to ThermoFisher, Microarray Research Services Laboratory at Santa Clara, CA, USA. A DNA concentration of 10ng/µl is required for genotyping. After applying this filter, *Pinus pinea*, *Pinus pinaster* and Populus sp did not reach the desired number of samples of 120 (respectively 2,5%, 14,2% and 14,8% of samples eliminated). In these species, the threshold was lowered to 9 ng/µl. We accepted this threshold to increase the number of samples but also to test the efficiency of genotyping with lower concentrations. This will be useful for the large genotyping that will be done for the tasks 2.2 and 3.2. For *Fraxinus* sp, the number of samples >10 ng/µl was still too high and 31 samples were discarded. In total, 108 *Pinus pinea*, 119 *Pinus pinaster*, 118 *Populus* sp and 135 *Fraxinus* sp will be used as genotyping panel to evaluate the screening array (Table 2, Annex 2).

Table 2. Number of SNP per species included in the screening array and number of individuals per species used as genotyping screening panel

Species	SNP	Genotypes
Pinus pinea	36 784	108
Pinus pinaster	137 321	119
Populus sp	137 545	118
Fraxinus excelsior	138 038	135
Total	449 688	480

4 Multiple species array 4TREE

4.1 Quality controls of the screening array

For each of the four species, all genotyped individuals were analyzed using Axiom Analysis Suite software (https://downloads.thermofisher.com/Axiom_Analysis_Suite_v_4.0.1_User_Guide.pdf). The analysis was performed with the recommended default parameters for diploid species. The data went through two steps of quality control. The first step concerned the quality control on the samples, which were filtered on Dish Quality Control > 0.82 (contrast of fluorescence signal, computed on monomorphic loci) and Quality Control Call Rate > 0.97 (call rate computed on high quality SNPs identified by ThermoFisher in calibration plates). The second step was the quality control on SNPs with thresholds characterizing the call rate and the position of the clusters compared to each other. Some of these parameters were adjusted from their default value in *Pinus sp*. and in *Fraxinus sp*. In *Pinus sp*. the call rate cut-off was lowered from 97% to 95%, as these species displayed less SNPs in the array, among which a large number were not recommended by ThermoFisher. In *Fraxinus sp*. the parameter HetSO (heterozygous strength offset) that characterize the position of the heterozygote cluster relative to the homozygote cluster was lowered from -0.1 to -0.3.

Adopting these metrics, the software classifies the SNP into categories (**Table 3**). Only SNPs from categories recommended by ThermoFisher were kept (PolyHighResolution, NoMinorHomozygous). The informative SNPs are included in the category PolyHighResolution where both homozygotes and the heterozygotes are represented. The NoMinorHomozygote ones, where there is only one of the homozygote cluster, were also considered for some markers of interest

Table 3. Genotype calling summary and success

		Pinus pinea	*Pinus pinaster*	*Fraxinus sp*	*Populus sp*
Samples	Total number of samples	**108**	**119**	**135**	**118**
	Passed samples	**108**	**111**	**126**	**115**

		Pinus pinea	*Pinus pinaster*	*Fraxinus sp*	*Populus sp*
	Av. cluster call rate	**99.65**	**99.48**	**99.51**	**99.58**
	Sample Reproducibility			**99.94**	**99.64**
SNPs	Total nb of SNPs	**36,784**	**133,408**	**138,038**	**137,545**
	PolyHighResolution (PHR)	**4,831**	**14,815**	**46,280**	**52,193**
	NoMinorHom (NMH)	**840**	**10,853**	**19,081**	**16,266**
	MonoHighResolution (MHR)	**20,329**	**50,260**	**27,651**	**3,407**
	Call rate below threshold (CRBT)	**880**	**5,450**	**8,107**	**5,987**
	Off target variant (OTV)	**518**	**3,755**	**7,557**	**18,998**
	Other	**9,386**	**48,275**	**29,326**	**40,694**
	% Genotyping success (PHR, NMH, MHR)	**70.68**	**56.91**	**67.38**	**52.25**
	% Informative SNPs (PHR, NMH)	**15.42**	**19.24**	**47.35**	**49.77**

Next paragraphs describe the filtering selection steps done on these SNPs to select on average the 12.5K SNPs per species in the final 4TREE array.

4.2 Selection of informative and robust SNP in Pinus pinea

The majority of the markers (20,329) were classified as MonoHighResolution (MHR) and thus resulted monomorphic in the analyzed dataset. Additional 10,784 markers did not pass the standard Axiom quality filtering, while the remaining 5,671 markers resulted polymorphic and with high quality. Of them, 4,831 were classified as PolyHighResolution, meaning that all the three possible genotypes have been observed in the dataset, while the remaining 840 markers were classified as NoMinorHom, meaning that the minor allele was never observed in homozygous state.

Due to the scarcity of available polymorphic markers, expected considering that *P. pinea* is classified as genetically depauperate species, all the 5,671 markers classified as PolyHighResolution or NoMinorHom have been selected to be included in the 4TREE array.

The number of markers selected for *P. pinea* was under the target number of 12,500 SNPs per species. That meant that 6,829 slots were available for the 3 other species, resulting in 2,300 additional slots available per species.

4.3 Selection of informative and robust SNP in Pinus pinaster

4.3.1 Performance of the Axiom screening array

Before proceeding with marker selection, we performed a quality check of the genotyping data obtained with the screening array for a set of approximately 4,000 markers that have been previously genotyped with the Illumina Infinium Array. In particular, we compared genotype calls obtained with both technologies (Illumina Infinium and ThermoFisher Axiom) in 70 different individuals. We obtained a very high percentage of concordant genotypes for the majority of the samples (**annex 3**), with a mean value of 98.9%. However, in eleven samples the levels of concordance were lower, with a mean value of about 53%. The quality of these samples was high and very similar to all the other samples so with high probability the low concordance observed for these 11 samples is caused by a mislabeling issue and not by a low performance of the genotyping.

4.3.2 Pinus pinaster markers selection

The majority of the markers (57,480) did not pass the standard Axiom quality filtering and additional 50,260 markers were classified as MonoHighResolution, thus resulting monomorphic in the analyzed dataset. The remaining 25,668 markers resulted polymorphic and showed high quality. Of them, 14,815 were classified as PolyHighResolution, while the remaining 10,853 were classified as NoMinorHom.

The 16,003 *P. pinaster* markers selected for the 4TREE array were classified into three priority levels:

- Priority 1: 3,889 markers classified as PolyHighResolution (3 441) or NoMinorHom (416) and already included in a previous Illumina Infinium array.
- Priority 2: 8,160 PolyHighResolution markers that are located at least 100 bp apart one from each other.
- Priority 3: 3,954 NoMinorHom that are located at least 100 bp apart from other selected markers.

4.4 Selection of highly informative SNP for Populus deltoides and P. nigra

4.4.1 Performance of the Axiom screening array

SNP selection for Populus species was performed with already available genomic resources developed for *P. nigra* (Faivre-Rampant et al. 2016; Rogier et al. 2018; Pégard et al. 2019) (GenTree project) and *P. deltoides* (Fahrenkrog et al. 2017). In this section, we compared genotype calling from former data and the Axiom screening array genotype calling for PolyHighResolution SNPs. To do so, SNP calling from DNA seq and RNA seq data and from the Infinium BlackPoplar array for several samples that where included in the *Populus* sp Axiom screening panel was available.

4.4.2 Concordance with DNA seq and RNA seq data

The Axiom screening array genotype calling of 23 individuals already genotyped in the previous studies was compared. The result of the comparison for each individual is reported in **annex 4**. Genotype concordance for 20 genotypes was > 90% (between 91 and 97%). Even so, genomic discordance was > 56% for Alaric, suggesting that this is probably not the same individual. More discordance than expected for the *P. nigra* individuals 71077-308 (86.96%) and *P. deltoides* V5 (80%) was observed, which can be explained by the lower quality of the DNAseq and exome capture data of these individuals.

4.4.3 Concordance across Infinium Black poplar array and Axiom screening array

We also compared the genotype calling on the 6,737 SNPs included in the Infinium BlackPoplar array and the Axiom screening array for 28 individuals. Genotype concordance, reported in **annex 5**, is high (superior to 97%). These comparisons showed a good concordance between our screening array PHR SNPs and existing data, indicating the robustness of these selected SNPs.

4.4.4 test of Mendelian segregation conformity

In the genotyping panel, 5 trios were included (2 parents and one progeny), 17 progenies and the male or female parent of the cross (the genotype of the other parent coming from previous studies). In total, there were 25 parent-descendant trios. In these trios, Mendelian segregation was tested. A trio is considered as coherent on one SNP if the genotype of the descendant can be obtained with the parent's genotypes.

The results of this test are displayed in **annex 6**:

- The *P. nigra* x *P. nigra* progenies are coherent with their parent at 98%.
- The *P. deltoides* x *P. nigra* hybrids are less coherent with 6 to 14% of incoherent SNPs. This could be partially explained assuming the presence of a null allele in one of the parent that could not be detected by the screening array. In that case, the hybrid could be heterozygote with one allele from the other parent and the null allele being incorrectly classified as homozygote.
- One *P. deltoides* x *P. deltoides* trio had as much incoherence as the *P. nigra* trios, while the two others are as incoherent as the *P. deltoides* x *P. nigra* trios.

4.4.5 Populus sp. SNP selection

The poplar SNPs belong to three different categories:

- Substitutions: each species is homozygote while the hybrids between them are heterozygote. They discriminate between the species (hereafter named "Substitutions").
- SNPs shared between *P. nigra* and *P. deltoides* (named "Shared").
- SNPs specific to *P. deltoides* (named "P.deltoides")
- SNPs specific to *P. nigra*, where 3 subcategories can be identified:
    - SNPs from the DNAseq and RNASeq data (named "P. nigra")

- SNP from the Infinium Back poplar (named “Illumina array”)
- SNPs from GWAS studies related to wood properties (named “GWAS”)

These SNPs went first through a series of filters, and then through selection. The first filter was to make sure that the data matched the category of the SNPs, meaning that some SNPs were re-categorized. For example, because the data showed they were not specific to a species. The second filter was to eliminate SNPs that were incoherent in the Mendelian segregation test. The SNPs were then submitted to selection. The main goal was to select SNPs well distributed along the chromosomes in each species while conserving the SNPs of interest. These processes are shown on **figures 1** and **2**.

In total 15,754 SNPs were selected, including 14,685 priority 1 SNPs and 1,069 additional priority 2 SNPs.

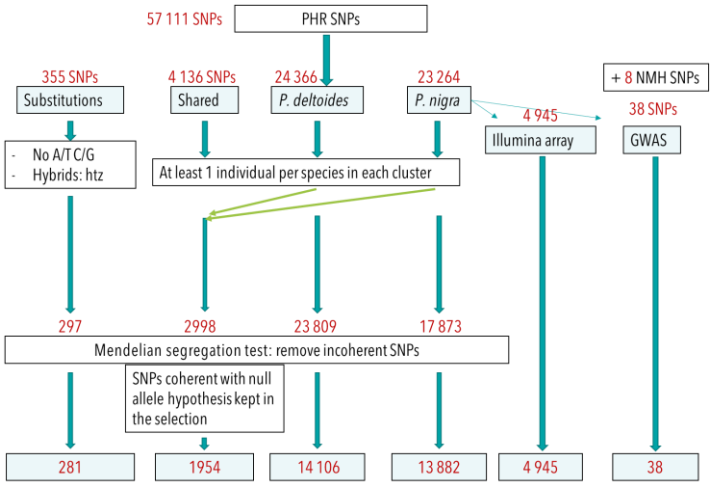


Figure 1. Filtering on *Populus* sp. Data.

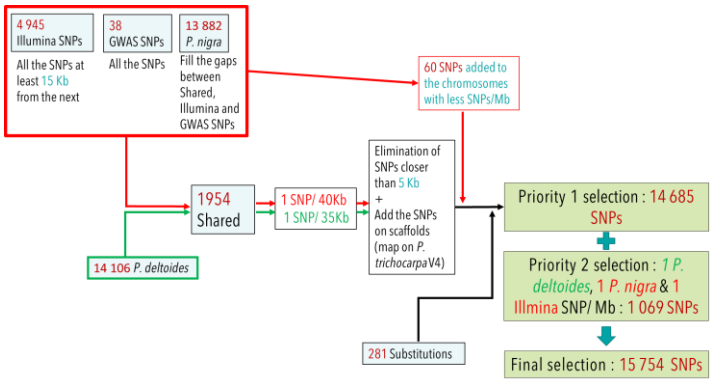


Figure 2. *Populus* sp. selection process

4.5 Selection of informative and robust SNP in Fraxinus sp

One of the main priorities for the *Fraxinus* species was to integrate into the array SNPs associated with characters of interest (ash borer and dieback disease resistance and incompatibility). Among these SNPs, 1,704 were placed into priority 1. This includes 20 SNPs from contigs associated with incompatibility, 51 associated with ash borer disease, and 1,633 with dieback disease. Because of their importance, these SNPs were selected among the PolyHighResolution, MonoHighResolution and NoMinorHom.

The priority 2 SNPs were also taken from these three categories of SNPs among those scored as "recommended". These 200 SNPs were selected to discriminate between two *Fraxinus* species: *F. angustifolia* and *F. excelsior*. They were chosen as the markers with the highest loadings on the first axis of a PCA analysis that discriminated all *F. excelsior* samples in the test panel from five pure *F. angustifolia* samples.

The other priorities were taken among the PolyHighResolution SNPs and focus on even coverage of *the F. excelsior* genome. The priority 3 SNPs consist of 9,698 SNPs, 1 SNP per contig. The priority 4 SNPs were 5,922 additional SNPs, one per contig. In total 17, 524 SNPs were selected for *Fraxinus sp*.

4.6 Summary for 4TREE array

A dedicated paper is under preparation for the elaboration of the 4TREE array.

Table 4. Summary of the SNP selection for 4TREE array

	Populus sp.	Pinus pinea	Pinus pinaster	Fraxinus sp.	Total
Priority 1	14,685	5,671	3,889	1,704	**25,949**
Priority 2	1,069		8,160	200	**9,429**
Priority 3			3,954	9,698	**13,652**
Priority 4				5,922	**5,922**
Total	**15,754**	**5,671**	**16,003**	**17,524**	**54,952**

A total of 54,952 SNPs was sent to ThermoFisher. The final array was composed of 45,893 selected to balance the number of SNPs per species and avoid A/T and C/G polymorphisms that take twice the place on the array. The array SNPs content was as detailed in **table 5**.

Table 5. 4TREE array composition

Populus sp.	*Pinus pinea*	*Pinus pinaster*	*Fraxinus* sp.	Total
13,407	5,671	13,407	13,407	45,892

5 Selection of informative and robust SNPs in *Picea abies*

In Norway spruce, three steps were used to construct the 50K SNP array. Firstly, we constructed a 450K pilot screening array based on 29 whole genome re-sequenced Norway spruce samples. Secondly, we screened the 450K SNPs based on 480 samples from species range-wide distribution and also Swedish breeding populations. Finally, to evaluate the final 50K array, we genotyped three sets of samples. The detailed description was already published (Bernhardsson et al. 2021). Here, we will concisely summarize the three steps.

5.1 Construction of the 450K pilot screening array

In order to design the 450K pilot array probes, 29 whole genome re-sequenced Norway spruce samples with a genome coverage > 15x were used for probe selection (Bernhardsson et al. 2021). After evaluation, all recommended intragenic SNPs (259,994) and the best ranked intergenic SNPs (190,499) were chosen for the 450K pilot array.

5.2 Selection of the 50K SNP array from the pilot screening array

For Norway spruce, we only select the SNPs from PHR category to the final 50K array, since there were enough candidates in the category. Nevertheless, we followed severe three filtering steps as indicated in **Figure 3** below. The detailed description has been described (Bernhardsson et al. 2021). Here, we described a few specific strategies used in maker selection of our Norway spruce. We attempted to cover as many of the genomic regions as possible by first selecting one SNP per scaffold. If an intragenic SNP within the scaffold was available, that SNP got prioritized, otherwise an intergenic SNP was randomly selected. Meanwhile, G/C and A/T SNPs were avoided whenever possible. To tag as many unique gene-models as possible, an additional 160 SNPs were selected to incorporate 160 gene-models not yet covered under the preceding procedure. We also included an additional 125 SNPs that were flanking known published associations (Baison et al. 2019; Elfstrand et al. 2020) or preliminary associations from GWAS on bud flush, bud set and wood quality traits (unpublished). Finally, an additional 1,608 SNPs were randomly selected to bring the total number of selected SNPs up to 47,445 which could fit on the 50K Axiom array together with ~ 2,000 control probes to account for background noise during imaging analysis.

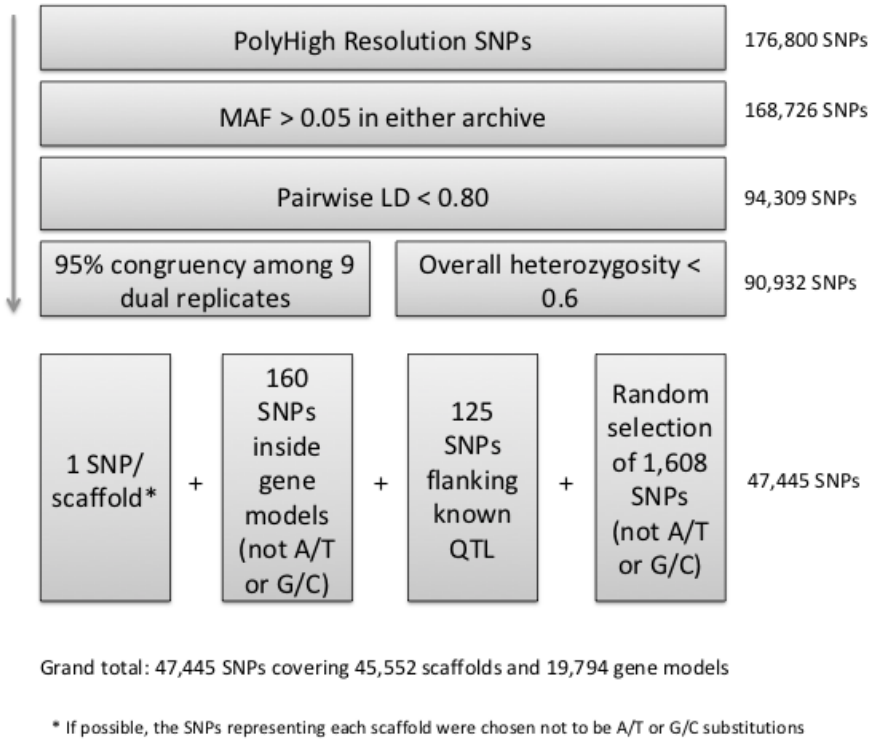


Figure 3. A schematic illustration of the probe selection pipeline from the 450K screening array to the final 50 K array.

5.3 Evaluation and validation of the 50K genotyping array

To evaluate the performance of the 50K genotyping array, we selected and genotyped another three sets of samples. First, four full sib Norway spruce families consisting of two parents and between 12 to 14 offspring were genotyped to estimate the Mendelian inheritance (MI) error rate. The MI error rate was calculated as the proportion of family trios that violate Mendelian inheritance rule. Second, 28 haploid megagametophytes were genotyped to evaluate the probe specificity and

examine whether probes were binding to different paralogues. For a 100% probe specificity, all genotyped megagametophytes should be homozygous. Therefore, a specificity error rate was calculated for each probe as the proportion of megagametophytes showing a heterozygous call. Third, 160 samples from three other spruce species (White spruce, Black spruce and Sitka spruce) were genotyped to evaluate the transferability of our spruce array to other spruce species.

6 Development of a species-specific genotyping array for *Pinus sylvestris* from screening array to final PiSy50k array

Genotyping array for *Pinus sylvestris* was developed using SNPs from eight different sources (Figure 3) including: exome captures (J. S. Tyrmi et al. 2019; J. Tyrmi et al. 2020) (Kesälahti et al. In prep), transcriptomes (Wachowiak et al. 2015; Ojeda et al. 2019), previous *Pinus sylvestris* Axiom PineGAP SNP-array (Perry et al. 2020), and candidate gene sets (Avia et al. 2014; Kujala et al. 2017; Vuosku et al. 2018; 2019). The main emphases were on selecting highly polymorphic and independent SNPs, avoiding paralogous regions of the genome, and including SNPs in candidate genes and on the previous Axiom PineGAP array. First, a set of 407,540 SNPs was screened by genotyping 480 samples from a Finnish natural population (Punkaharju ISS) on a screening array, and then a set of 47,712 consistently high-quality SNPs were selected for the final PiSy50k array.

When choosing SNPs for the PiSy50k array based on the screening array, the polymorphic conversion types PHR and NMH with call rate (CR) >= 0.97 were considered successful. Altogether 245 149 SNPs (60.2%) out of 407 540 converted successfully in the screening array of which 157 325 were polymorphic. The highest priority was given to SNPs from candidate gene sources (Noveltree and LUKE PacBio) and the previous Axiom PineGAP array (UKCEH1). After these SNPs had been included, the rest of the array was filled with SNPS from the other SNP sources. Somewhat differing thresholds were used for SNPs from different sources, but all SNPs were chosen to show: high minor allele frequency (MAF), high independence (low linkage disequilibrium, LD), no indication of potentially paralogous regions (Hardy-Weinberg equilibrium and no/low heterozygosity in haploid samples), and a low number of genotyping errors (Mendelian inheritance errors; Figure 4).

Of the eight data sources, markers from RNA-seq origin were the most prevalent (20,797 SNPs, 44%, Table 4). Also, a considerable number (6718, 14%) of SNPs on Axiom PineGAP array were included in the PiSy50k array.

Table 4. Number and proportions of markers from each source at different steps of the PiSy50k array design (Kastally et al. in Prep).

Data set ID	Initial set		Screening array		PiSy50k array		From the Screening to the PiSy50k array
	count	%	count	%	count	%	%
UKCEH1	20795	0.5	18901	4.6	6718	14.1	35.5
UKCEH2	175841	4.6	57261	14.1	7788	16.3	13.6
LUKE PacBio	6157	0.2	5980	1.5	531	1.1	8.9
NovelTree	3584	0.1	3273	0.8	1187	2.5	36.3
ProCoGen exome diploid	304661	8	23565	5.8	340	0.7	1.4
ProCoGen exome haploid	1870598	48.9	94180	23.1	6995	14.7	7.4
UOULU exomeFEB2019	95504	2.5	25140	6.2	3356	7	13.3
UOULU RNAseq	1349291	35.3	179240	44	20797	43.6	11.6
Total	**3826431**	**100**	**407540**	**100**	**47712**	**100**	**11.7**

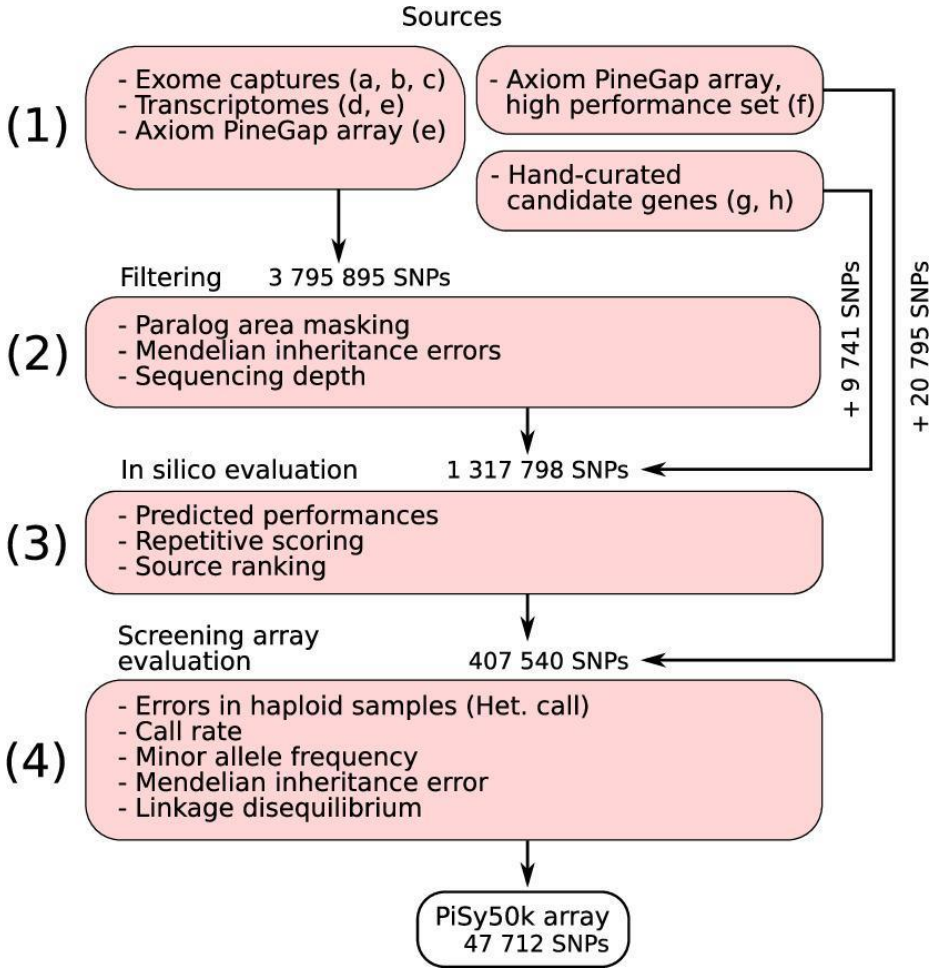


Figure 4. Workflow of the development of PiSy50k SNP array for Pinus sylvestris (Kastally et al. In Prep). The selection was done in four steps: (1) the collection of SNPs from 8 sources (a: ProCoGen exomeHaploid, b: ProCoGen exomeDiploid, c: UOULU exomeFEB2019, d: UOULU RNAseq, e: UKCEH2, f: UKCEH1, g: NovelTree and h: LUKE PacBio); (2) a filtering step removing potential errors from paralogous genomic areas; (3) an evaluation by Thermo Fisher Scientific to retain the best set of 407 540 markers (Screening set); (4) a filtering step based on the Screening set performance to retain the best 47 712 markers retained in the PiSy50k array.

7 Conclusion

The screening array strategy was a good way to test the SNPs for the final arrays. We observed genotyping success between 52 and 70%. This result was expected and coherent with what was previously observed in Axiom screening arrays in the literature, for example 54% in *Pyrus* (Montanari et al. 2019), and in the other arrays developed in B4EST. In the Scots pine array, genotyping success was 60% (T.Pyhajarvi, personal communication). A similar rate was obtained in Norway spruce with 57.6% of PHR, NMH and MHR SNPs (H.Wu, personal communication).

With each species being able to extract the best from the screening array and put forward SNPs of interest we can expect the final product to be composed of robust SNPs and have a high success rate.

To conclude, a total of 54,952 SNPs were sent to Thermofisher Bioinformatic service to select the best combination in order to build the final product array named Axiom_4tree array. A total of 45,893 SNPs were selected for the array, including 5,671 *P.pinea* SNPs, 13,407 *P.pinaster* SNPs, 13,407 *Populus sp*. SNPs and 13,407 *Fraxinus sp*. SNPs. For the *P. sylvestris* array, PiSy50K array, the amount was 47,712 SNPS. For *Picea abies* array, the number od total SNPs was 47,445.

8 Partners involved in the work

Organism	*Country*	*People involved*	*Species*
CNR: Institute of Biosciences and Bioresources, National Research Council	*Italy*	*S. Pinosio* *I. Spanu* *G.G. Vendramin*	*Pinus pinea* *Pinus pinaster*
CREA: Council for Agricultural Research and Economics	Italy	*C. Biselli,* *L. Cattivelli* *A. Fricano* *G. Nervo*	*Populus sp.*
INRA: National Institute of Agronomic Research	*France*	*A. Dowkiw* *P. Faivre Rampant* *V. Guerin* *S.C. González-Martínez* *R. Guilbaud* *V. Jorge* *O. Rogier* *L. Sanchez Rodriguez (WP3 leader)* *V. Segura*	*Populus sp.* *Pinus pinea* *Pinus pinaster* *Fraxinus sp.*
*WUR: Wageningen University and Research**	*Netherlands*	*J. Buiteveld* *P. Copini* *D. Esselink* *L. Kodde* *M.J.M. Smulders*	*Fraxinus sp.*

** With the participation of R.Buggs team : Royal Botanic Gardens Kew, Richmond, Surrey TW9 3AB, UK & School of Biological and Chemical Sciences, Queen Mary University of London, London, E1 4NS, U*

9 Annexes

Annex 1: List of SNP arrays in forest trees

Author	Title	Species	Type of array	Number of markers (bp)	SNP origin	Uses
Azaiez et al. 2018	A catalog of annotated high-confidence SNPs from exome capture and sequencing reveals highly polymorphic genes in Norway spruce (*Picea abies*)	*Picea abies*	Illumina Infinium	5 571	Exome capture	Comparative genome mapping, association mapping, landscape genomic studies
Ćalić et al. 2017	Genome-wide association study identifies a major gene for beech bark disease resistance in American beech (*Fagus grandifolia* Ehrh.)	*Fagus grandifolia*	Affymetrix Axiom	5 838	Transcriptome assembly	GWAS
Chancerel et al. 2011	Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine	*Pinus pinaster*	Illumina GoldenGate	1 536	DNAseq on candidate genes Sanger-derived Expressed Sequenced Tags	comparative genomics among the conifers
Eckert et al. 2009	High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.)	*Pinus taeda*	Illumina GoldenGate	384	DNAseq of candidate genes	Genetic mapping, associations studies

Author	Title	Species	Type of array	Number of markers (bp)	SNP origin	Uses
Faivre-Rampant et al. 2016	New resources for genetic studies in *Populus nigra* : genome-wide SNP discovery and development of a 12k Infinium array	*Populus nigra*	Illumina Infinium	12 000	Whole genome resequencing	Population genetics, association genetics (QTL identification)
Geraldes et al. 2013	A 34K SNP genotyping array for *Populus trichocarpa* : Design, application to the study of natural populations and transferability to other *Populus* species	*Populus trichocarpa*	Illumina Infinium	34 000	DNAseq and RNAseq on candidate genes	Study of the genetic architecture of complex trait variation (ie association study-QTL identification) Study of the differentiation between populations
Grattapaglia et al. 2011	High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species	*Eucalyptus sp.*	Illumina GoldenGate	768	Sanger-derived ESTs, candidate SNPs	genome-wide phylogenetic, population genomics studies, genomic selection
Howe et al. 2013	A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation	*Pseudotsuga menziesii*	Illumina Infinium	8 067	cDNA sequencing/ transcriptome sequencing	Genomic selection, genome x environment interactions
Kubisiak et al. 2013	A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*)	*Castanea mollissima*	Illumina GoldenGate	1 536	EST	Linkage mapping

Author	Title	Species	Type of array	Number of markers (bp)	SNP origin	Uses
Lepoittevin et al. 2015	Single-nucleotide polymorphism discovery and validation in high-density SNP array for genetic analysis in European white oaks	*Quercus sp.*	Illumina Infinium	7 913	ESTs, in silico SNPs from previous studies	Comparative studies between white oak species
Lepoittevin et al. 2010	In Vitro vs In Silico Detected SNPs for the Development of a Genotyping Array: What Can We Learn from a Non-Model Species?	*Pinus pinaster*	Illumina GoldenGate	384	re-sequenced candidate genes (in vitro) + ESTs (in silico)	Building consensus map, association mapping
Marrano et al. 2019	A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density Axiom™ *J. regia* 700K SNP genotyping array	*Juglans regia*	Affymetrix Axiom	700 000	Whole genome sequencing	Association genetics (GWAS), genomic selection (GS)
Pavy et al. 2016	Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (*Picea mariana*)	*Picea mariana*	Illumina Infinium	97 075	Exome capture	landscape genomics, association studies, genomic selection, genetic certification
Pavy et al. 2013	Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners	*Picea glauca*	Illumina Infinium	7 338 and 9 559	Gene resequencing and EST	Population genetics and genetic mapping

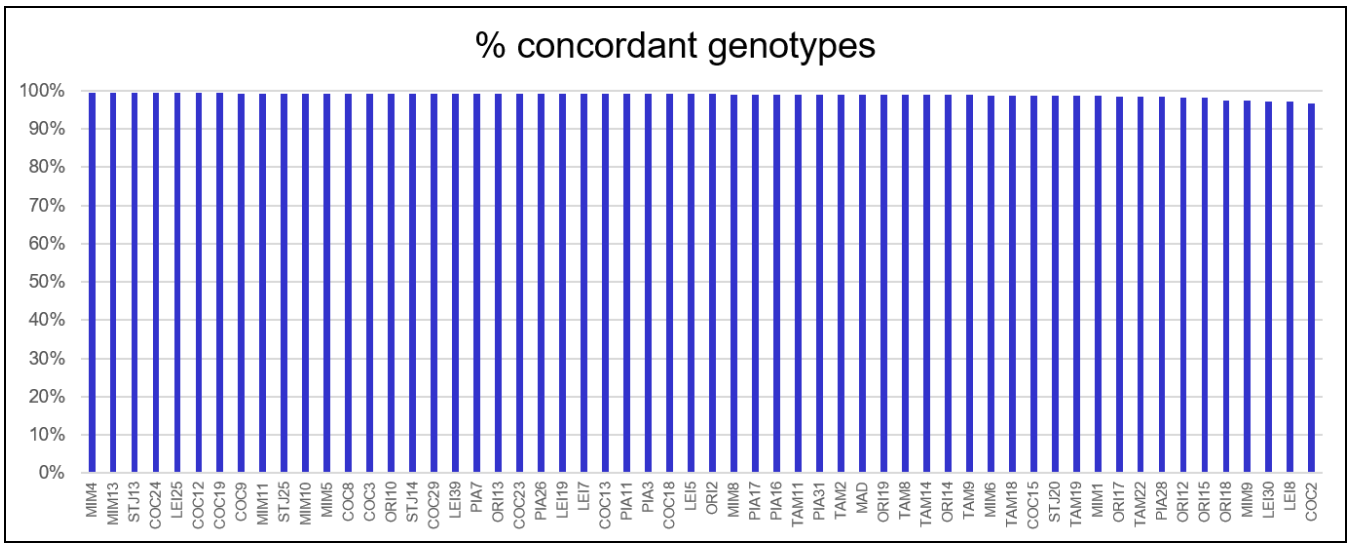
Author	Title	Species	Type of array	Number of markers (bp)	SNP origin	Uses
Plomion et al. 2016	High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*)	*Pinus pinaster*	Illumina Infinium	9 000	Previously available SNPs + EST	Population genetics, conservation genetics, breeding
Silva-Junior et al. 2015	A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species	*Eucalyptus sp.*	Illumina Infinium	60 000	Whole genome resequencig	Genomic selection, GWAS
Uchiyama et al. 2012	Single nucleotide polymorphisms in *Cryptomeria japonica*: their discovery and validation for genome mapping and diversity studies	*Cryptomeria japonica*	Illumina GoldenGate	1 536	EST re-sequencing	QTL studies, GS, marker assisted selection
Wong et al. 2012	Development of high-throughput SNP-based genotyping in *Acacia auriculiformis* x *A. mangium* hybrids using short-read transcriptome data	*Acacia sp.*	Illumina GoldenGate	768 genotyped (96 validated)	Transcriptome assembly	Linkage map, comparative genomics

Annex 2. Number of selected tree per species and their geographical origin sent to ThermoFisher for validation of the screening array

Species	Location	Number of individuals
Pinus pinea	Italy	6
	Lebanon	8
	Turkey	8
	France	6
	Spain	57
	Portugal	35
Pinus pinaster	France	76
	Morocco	11
	Portugal	10
	Spain	23
Populus deltoides	Illinois	11
	Iowa	3
	Kansas	2
	Kentucky	1
	Louisiana	1
	Michigan	2
	Minnesota	3
	Mississippi	4
	Missouri	2
	Nebraska	2
	Ohio	3
	Oklahoma	5
	Tennessee	2
	Texas	2
	Winsconsin	2
***P. deltoides* progenies**		4
Populus nigra	Belgium	1
	France	16 (+1 duplicate)
	Great Britain	1
	Germany	5
	Greece	1
	Hungary	1
	Ireland	1
	Italy	19 (+1 duplicate)
	Netherlands	5
	Portugal	2
	Spain	2
	Turkey	1
	Yugoslavia	1
***P.nigra* progenies**		11
P.deltoides x P.nigra		12
Fraxinus angustifolia	France	13
	Unknown	1

Species	Location	Number of individuals
Fraxinus excelsior	Denmark	2
	France	44
	Ireland	2
	Italy	2
	Lithuania	2
	Netherlands	86
	Poland	2
Fraxinus mandshurica	France	3
Fraxinus sp.	Unknown	13
***F.excelsior* progenies**		7

Annex 3. Concordance between Axiom and Infinium arrays genotypes *in P. pinaster*.



Annex 4. Concordance with DNA seq and RNA seq data *for Populus sp.*

Individual	**% concordance**	**Number of shared SNPs (PHR)**	**Data origin**	**Species**
73028-62	**93.87**	9,687	WGS	*P. deltoides*
ALARIC	**43.44**	9,765	WGS	*P. deltoides*
L150-089	**93.98**	9,719	WGS	*P. deltoides*
V5	**80.00**	485	Exome capture	*P. deltoides*
662200037	**94.04**	25,636	WGS	*P. nigra*
662200039	**94.84**	25,899	WGS	*P. nigra*
662200057	**95.10**	26,016	WGS	*P. nigra*
662200098	**96.03**	26,248	WGS	*P. nigra*
662200357	**91.70**	25,266	WGS	*P. nigra*
662200495	**96.17**	26,299	WGS	*P. nigra*
662201338	**94.66**	26,112	WGS	*P. nigra*
3824-3	**95.88**	26,017	WGS	*P. nigra*
71041-3-402	**93.57**	25,487	WGS	*P. nigra*
71069-914	**96.64**	26,344	WGS	*P. nigra*
71077-308	**86.96**	25,543	WGS	*P. nigra*
Blanc de Garonne	**96.44**	26,400	WGS	*P. nigra*
DRA-13	**94.17**	11,476	RNAseq	*P. nigra*
H480	**94.66**	25,969	WGS	*P. nigra*
KUH-40	**93.96**	11,526	RNAseq	*P. nigra*
NL-1171	**93.18**	11,309	RNAseq	*P. nigra*
NL-1322	**94.20**	11,395	RNAseq	*P. nigra*
SAN_GIORGIO	**96.11**	26,121	WGS	*P. nigra*
SAN_GIORGIO	**97.58**	2,187	Exome capture	*P. nigra*
SPM-13	**95.99**	2,168	RNAseq	*P. nigra*

Annex 5. Concordance with Infinium Black Poplar array data

Individual	% concordance	Nb of SNPs	Species
662200001	**98,22**	4541	*P. nigra*
662200037	**98,72**	4540	*P. nigra*
662200039	**98,46**	4539	*P. nigra*
662200057	**98,30**	4525	*P. nigra*
662200098	**98,61**	4540	*P. nigra*
662200203	**98,61**	4539	*P. nigra*
662200311	**98,63**	4535	*P. nigra*
662200357	**98,48**	4535	*P. nigra*
662200457	**98,33**	4538	*P. nigra*
662200495	**98,32**	4530	*P. nigra*
662201338	**97,59**	4532	*P. nigra*
3824-3	**97,64**	4530	*P. nigra*
71041-3-402	**99,31**	4935	*P. nigra*
71069-914	**98,35**	4535	*P. nigra*
71077-308	**98,99**	4934	*P. nigra*
DRA-01	**99,23**	4936	*P. nigra*
DRA-13	**98,90**	4902	*P. nigra*
DRA-21	**99,29**	4940	*P. nigra*
DRA-35	**98,99**	4936	*P. nigra*
H480	**97,02**	4534	*P. nigra*
KUH-40	**98,99**	4934	*P. nigra*
KUH-5	**99,27**	4933	*P. nigra*
KUH-9	**99,17**	4938	*P. nigra*
NL-1171	**98,69**	4826	*P. nigra*
NL-1322	**98,66**	4856	*P. nigra*
NL-1324	**99,21**	4940	*P. nigra*
SPM-12	**99,39**	4941	*P. nigra*
SPM-13	**99,17**	4938	*P. nigra*
VDL-07	**99,19**	4933	*P. nigra*

Annex 6. Coherence of Mendelian segregation in trios in *Populus nigra*.

Trio	Nb of SNPs	% Coherent	Cross
L155-079_73021-15_661101717	9,712	**84.10**	*P. deltoides* x *P.deltoides*
L155-079_MSS050-27_661101878	9,603	**96.88**	*P. deltoides* x *P.deltoides*
TNS060-042_73021-15_661101916	51,898	**84.17**	*P. deltoides* x *P.deltoides*
L155-079_Blanc de Garonne_661200506	9,548	**92.42**	*P. deltoides* x *P.nigra*
L155-079_Blanc de Garonne_661200531	9,674	**93.60**	*P. deltoides* x *P.nigra*
L155-079_Blanc de Garonne_661200533	9,660	**93.51**	*P. deltoides* x *P.nigra*
L155-079_Blanc de Garonne_661200538	9,677	**93.58**	*P. deltoides* x *P.nigra*
L155-079_Blanc de Garonne_661200539	9,672	**93.64**	*P. deltoides* x *P.nigra*
L155-079_Blanc de Garonne_661200552	9,700	**93.82**	*P. deltoides* x *P.nigra*
L155-079_Blanc de Garonne_661200558	9,676	**93.70**	*P. deltoides* x *P.nigra*
L155-079_71077-308_661200592	1,217	**91.04**	*P. deltoides* x *P.nigra*
TNS060-042_Blanc de Garonne_661200701	50,844	**86.08**	*P. deltoides* x *P.nigra*
TNS060-042_Blanc de Garonne_661200705	51,595	**85.78**	*P. deltoides* x *P.nigra*
TNS060-042_Blanc de Garonne_661200734	51,447	**85.69**	*P. deltoides* x *P.nigra*
VGN-CZB25_Blanc de Garonne_662200001	26,154	**88.59**	*P. nigra* x *P.nigra*
VGN-CZB25_Blanc de Garonne_662200037	26,179	**98.37**	*P. nigra* x *P.nigra*
VGN-CZB25_Blanc de Garonne_662200039	26,183	**98.34**	*P. nigra* x *P.nigra*
VGN-CZB25_Blanc de Garonne_662200057	26,070	**98.47**	*P. nigra* x *P.nigra*
VGN-CZB25_71077-308_662200098	25,307	**98.44**	*P. nigra* x *P.nigra*
VGN-CZB25_Blanc de Garonne_662200203	26,181	**98.38**	*P. nigra* x *P.nigra*
71041-3-402_Blanc de Garonne_662200311	51,817	**99.28**	*P. nigra* x *P.nigra*
71041-3-402_71077-308_662200357	25,495	**98.45**	*P. nigra* x *P.nigra*
71072-501_Blanc de Garonne_662200457	26,117	**98.44**	*P. nigra* x *P.nigra*
71072-501_71077-308_662200495	25,259	**98.33**	*P. nigra* x *P.nigra*
662200037_72131-017_662201338	25,787	**97.98**	*P. nigra* x *P.nigra*

10 Bibliography

Alonso, Pablo J., Millán Cortizo, Francisco R. Cantón, Belén Fernández, Ana Rodríguez, María Luz Centeno, Francisco M. Cánovas, and Ricardo J. Ordás. 2007. 'Identification of Genes Differentially Expressed during Adventitious Shoot Induction in Pinus Pinea Cotyledons by Subtractive Hybridization and Quantitative PCR.' *Tree Physiology* 27 (12): 1721–30. https://doi.org/10.1093/treephys/27.12.1721.

Azaiez, Aïda, Nathalie Pavy, Sébastien Gérardi, Jérôme Laroche, Brian Boyle, France Gagnon, Marie-Josée Mottet, Jean Beaulieu, and Jean Bousquet. 2018. 'A Catalog of Annotated High-Confidence SNPs from Exome Capture and Sequencing Reveals Highly Polymorphic Genes in Norway Spruce (Picea Abies)'. *BMC Genomics* 19 (1): 942. https://doi.org/10.1186/s12864-018-5247-z.

Avia, Komlan, Katri Kärkkäinen, Ulf Lagercrantz, and Outi Savolainen. 2014. 'Association of FLOWERING LOCUS T/TERMINAL FLOWER 1-like Gene FTL2 Expression with Growth Rhythm in Scots Pine (*Pinus sylvestris*)'. *New Phytologist* 204 (1): 159–70. https://doi.org/10.1111/nph.12901.

Baison, John, Amaryllis Vidalis, Linghua Zhou, Zhi-Qiang Chen, Zitong Li, Mikko J. Sillanpää, Carolina Bernhardsson, et al. 2019. 'Genome-wide Association Study Identified Novel Candidate Loci Affecting Wood Formation in Norway Spruce'. *The Plant Journal* 100 (1): 83–100. https://doi.org/10.1111/tpj.14429.

Bernhardsson, Carolina, Yanjun Zan, Zhiqiang Chen, Pär K. Ingvarsson, and Harry X. Wu. 2021. 'Development of a Highly Efficient 50K Single Nucleotide Polymorphism Genotyping Array for the Large and Complex Genome of Norway Spruce (*Picea Abies* L. Karst) by Whole Genome Resequencing and Its Transferability to Other Spruce Species'. *Molecular Ecology Resources* 21 (3): 880–96. https://doi.org/10.1111/1755-0998.13292.

Ćalić, Irina, Jennifer Koch, David Carey, Charles Addo-Quaye, John E. Carlson, and David B. Neale. 2017. 'Genome-Wide Association Study Identifies a Major Gene for Beech Bark Disease Resistance in American Beech (Fagus Grandifolia</Em> Ehrh.)'. *BMC Genomics* 18 (547): 1–14. https://doi.org/10.1186/s12864-017-3931-z.

Chancerel, Emilie, Camille Lepoittevin, Grégoire Le Provost, Yao-Cheng Lin, Juan Pablo Jaramillo-Correa, Andrew J Eckert, Jill L Wegrzyn, et al. 2011. 'Development and Implementation of a Highly-Multiplexed SNP Array for Genetic Mapping in Maritime Pine and Comparative Mapping with Loblolly Pine'. *BMC Genomics* 12 (1): 368. https://doi.org/10.1186/1471-2164-12-368.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. 'A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data'. *Nature Genetics* 43 (5): 491–98. https://doi.org/10.1038/ng.806.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. 'STAR: Ultrafast

Universal RNA-Seq Aligner'. *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Elfstrand, Malin, John Baison, Karl Lundén, Linghua Zhou, Ingrid Vos, Hernan Dario Capador, Matilda Stein Åslund, et al. 2020. 'Association Genetics Identifies a Specifically Regulated Norway Spruce Laccase Gene, PALAC5 , Linked to *Heterobasidion Parviporum* Resistance'. *Plant, Cell & Environment* 43 (7): 1779–91. https://doi.org/10.1111/pce.13768.

Eckert, Andrew J., Barnaly Pande, Elhan S. Ersoz, Mark H. Wright, Vanessa K. Rashbrook, Charles M. Nicolet, and David B. Neale. 2009. 'High-Throughput Genotyping and Mapping of Single Nucleotide Polymorphisms in Loblolly Pine (Pinus Taeda L.)'. *Tree Genetics & Genomes* 5 (1): 225–34. https://doi.org/10.1007/s11295-008-0183-8.

Fahrenkrog, Annette M., Leandro G. Neves, Márcio F. R. Resende, Christopher Dervinis, Ruth Davenport, W. Brad Barbazuk, and Matias Kirst. 2017. 'Population Genomics of the Eastern Cottonwood (Populus Deltoides)'. *Ecology and Evolution* 7 (22): 9426–40. https://doi.org/10.1002/ece3.3466.

Faivre-Rampant, P., G. Zaina, V. Jorge, S. Giacomello, V. Segura, S. Scalabrin, V. Guérin, et al. 2016. 'New Resources for Genetic Studies in Populus Nigra : Genome-Wide SNP Discovery and Development of a 12k Infinium Array'. *Molecular Ecology Resources* 16 (4): 1023–36. https://doi.org/10.1111/1755-0998.12513.

Geraldes, A., S. P. DiFazio, G. T. Slavov, P. Ranjan, W. Muchero, J. Hannemann, L. E. Gunter, et al. 2013. 'A 34K SNP Genotyping Array for Populus Trichocarpa : Design, Application to the Study of Natural Populations and Transferability to Other Populus Species'. *Molecular Ecology Resources* 13 (2): 306–23. https://doi.org/10.1111/1755-0998.12056.

Grattapaglia, Dario, Orzenil B. Silva-Junior, Matias Kirst, Bruno Marco de Lima, Danielle A. Faria, and Georgios J. Pappas. 2011. 'High-Throughput SNP Genotyping in the Highly Heterozygous Genome of Eucalyptus: Assay Success, Polymorphism and Transferability across Species'. *BMC Plant Biology* 11 (1): 65. https://doi.org/10.1186/1471-2229-11-65.

Howe, Glenn T., Jianbin Yu, Brian Knaus, Richard Cronn, Scott Kolpak, Peter Dolan, W. Walter Lorenz, and Jeffrey FD Dean. 2013. 'A SNP Resource for Douglas-Fir: De Novotranscriptome Assembly and SNP Detection and Validation'. *BMC Genomics* 14 (1): 137. https://doi.org/10.1186/1471-2164-14-137.

Kelly, Laura J., William J. Plumb, David W. Carey, Mary E. Mason, Endymion D. Cooper, William Crowther, Alan T. Whittemore, Stephen J. Rossiter, Jennifer L. Koch, and Richard J. A. Buggs. 2019. 'Genes for Ash Tree Resistance to an Insect Pest Identified via Comparative Genomics'. *BioRxiv*, September, 772913. https://doi.org/10.1101/772913.

Kubisiak, TL, Charles Nelson, Margaret Staton, T. Zhebentyayeva, Chris Smith, Bode Olukolu, Fang G-C, et al. 2013. 'A Transcriptome-Based Genetic Map of Chinese Chestnut, (Castanea Mollissima), and Identification of Regions of Segmental Homology with Peach (Prunus Persica)'. *Tree Genetics & Genomes* 9 (April): 557–71. https://doi.org/10.1007/s11295-012-0579-3.

Kujala, S T, T Knürr, K Kärkkäinen, D B Neale, M J Sillanpää, and O Savolainen. 2017. 'Genetic Heterogeneity Underlying Variation in a Locally Adaptive Clinal Trait in *Pinus sylvestris* Revealed by a Bayesian Multipopulation Analysis'. *Heredity* 118 (5): 413–23. https://doi.org/10.1038/hdy.2016.115.

Lepoittevin, C., C. Bodénès, E. Chancerel, L. Villate, T. Lang, I. Lesur, C. Boury, et al. 2015. 'Single-Nucleotide Polymorphism Discovery and Validation in High-Density SNP Array for Genetic Analysis in European White Oaks'. *Molecular Ecology Resources* 15 (6): 1446–59. https://doi.org/10.1111/1755-0998.12407.

Lepoittevin, Camille, Jean-Marc Frigerio, Pauline Garnier-Géré, Franck Salin, María-Teresa Cervera, Barbara Vornam, Luc Harvengt, and Christophe Plomion. 2010. 'In Vitro vs In Silico Detected SNPs for the Development of a Genotyping Array: What Can We Learn from a Non-Model Species?' Edited by Pär K. Ingvarsson. *PLoS ONE* 5 (6): e11034. https://doi.org/10.1371/journal.pone.0011034.

Li, Heng. 2011. 'A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data'. *Bioinformatics* 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.

Marrano, Annarita, Pedro J. Martínez-García, Luca Bianco, Gina M. Sideli, Erica A. Di Pierro, Charles A. Leslie, Kristian A. Stevens, et al. 2019. 'A New Genomic Tool for Walnut (Juglans Regia L.): Development and Validation of the High-Density Axiom™ J. Regia 700K SNP Genotyping Array'. *Plant Biotechnology Journal* 17 (6): 1027–36. https://doi.org/10.1111/pbi.13034.

Montanari, Sara, Luca Bianco, Brian J. Allen, Pedro J. Martínez-García, Nahla V. Bassil, Joseph Postman, Mareike Knäbel, et al. 2019. 'Development of a Highly Efficient Axiom™ 70 K SNP Array for Pyrus and Evaluation for High-Density Mapping and Germplasm Characterization'. *BMC Genomics* 20 (1): 331. https://doi.org/10.1186/s12864-019-5712-3.

Mutke, Sven, Rafael Calama, Santiago C. González-Martínez, Gregorio Montero, F. Javier Gordo, David Bono, and Luis Gil. 2012. 'Mediterranean Stone Pine: Botany and Horticulture'. In *Horticultural Reviews*, edited by Jules Janick, 153–201. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118100592.ch4.

Ojeda, Dario I., Tiina M. Mattila, Tom Ruttink, Sonja T. Kujala, Katri Kärkkäinen, Jukka-Pekka Verta, and Tanja Pyhäjärvi. 2019. 'Utilization of Tissue Ploidy Level Variation in *de Novo* Transcriptome Assembly of *Pinus Sylvestris*'. *G3: Genes|Genomes|Genetics* 9 (10): 3409–21. https://doi.org/10.1534/g3.119.400357.

Pavy, Nathalie, France Gagnon, Astrid Deschênes, Brian Boyle, Jean Beaulieu, and Jean Bousquet. 2016. 'Development of Highly Reliable in Silico SNP Resource and Genotyping Assay from Exome Capture and Sequencing: An Example from Black Spruce (Picea Mariana)'. *Molecular Ecology Resources* 16 (2): 588–98. https://doi.org/10.1111/1755-0998.12468.

Pavy, Nathalie, France Gagnon, Philippe Rigault, Sylvie Blais, Astrid Deschênes, Brian Boyle, Betty Pelgas, et al. 2013. 'Development of High-Density SNP Genotyping Arrays for White Spruce (Picea Glauca) and Transferability to Subtropical and Nordic Congeners'. *Molecular Ecology Resources* 13 (2): 324–36. https://doi.org/10.1111/1755-0998.12062.

Pégard, Marie, Odile Rogier, Aurélie Bérard, Patricia Faivre-Rampant, Marie-Christine Le Paslier, Catherine Bastien, Véronique Jorge, and Leopoldo Sánchez. 2019. 'Sequence Imputation from Low Density Single Nucleotide Polymorphism Panel in a Black Poplar Breeding Population'. *BMC Genomics* 20 (1): 302. https://doi.org/10.1186/s12864-019-5660-y.

Perry, Annika, Witold Wachowiak, Alison Downing, Richard Talbot, and Stephen Cavers. 2020. 'Development of a Single Nucleotide Polymorphism Array for Population Genomic Studies in Four European Pine Species'. *Molecular Ecology Resources* 20 (6): 1697–1705. https://doi.org/10.1111/1755-0998.13223.

Plomion, C., J. Bartholomé, I. Lesur, C. Boury, I. Rodríguez-Quilón, H. Lagraulet, F. Ehrenmann, et al. 2016. 'High-Density SNP Assay Development for Genetic Analysis in Maritime Pine (Pinus Pinaster)'. *Molecular Ecology Resources* 16 (2): 574–87. https://doi.org/10.1111/1755-0998.12464.

Rasheed, Awais, Yuanfeng Hao, Xianchun Xia, Awais Khan, Yunbi Xu, Rajeev K. Varshney, and Zhonghu He. 2017. 'Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives'. *Molecular Plant* 10 (8): 1047–64. https://doi.org/10.1016/j.molp.2017.06.008.

Rogier, Odile, Aurélien Chateigner, Souhila Amanzougarene, Marie-Claude Lesage-Descauses, Sandrine Balzergue, Véronique Brunaud, José Caius, Ludivine Soubigou-Taconnat, Véronique Jorge, and Vincent Segura. 2018. 'Accuracy of RNAseq Based SNP Discovery and Genotyping in *Populus nigra*'. *BMC Genomics* 19 (1): 909. https://doi.org/10.1186/s12864-018-5239-z.

Scheben, Armin, Jacqueline Batley, and David Edwards. 2017. 'Genotyping-by-Sequencing Approaches to Characterize Crop Genomes: Choosing the Right Tool for the Right Application'. *Plant Biotechnology Journal* 15 (2): 149–61. https://doi.org/10.1111/pbi.12645.

Silva-Junior, Orzenil B., Danielle A. Faria, and Dario Grattapaglia. 2015. 'A Flexible Multi-Species Genome-Wide 60K SNP Chip Developed from Pooled Resequencing of 240 Eucalyptus Tree Genomes across 12 Species'. *New Phytologist* 206 (4): 1527–40. https://doi.org/10.1111/nph.13322.

Sollars, Elizabeth S. A., Andrea L. Harper, Laura J. Kelly, Christine M. Sambles, Ricardo H. Ramirez-Gonzalez, David Swarbreck, Gemy Kaithakottil, et al. 2017. 'Genome Sequence and Genetic Diversity of European Ash Trees'. *Nature* 541 (7636): 212–16. https://doi.org/10.1038/nature20786.

Stocks, Jonathan J., Carey L. Metheringham, William Plumb, Steve J. Lee, Laura J. Kelly, Richard A. Nichols, and Richard J. A. Buggs. 2019. 'Genomic Basis of European Ash Tree Resistance to Ash Dieback Fungus'. *BioRxiv*, May, 626234. https://doi.org/10.1101/626234.

Torkamaneh, Davoud, Brian Boyle, and François Belzile. 2018. 'Efficient Genome-Wide Genotyping Strategies and Data Integration in Crop Plants'. *Theoretical and Applied Genetics* 131 (3): 499–511. https://doi.org/10.1007/s00122-018-3056-z.

Tyrmi, Jaakko S., Jaana Vuosku, Juan J. Acosta, Zhen Li, Lieven Sterck, Outi Savolainen, and Tanja Pyhäjärvi. 2019. 'Genomics of Clinal Local Adaptation in *Pinus Sylvestris* under

Continuous Environmental and Spatial Genetic Setting'. *BioRxiv*, May, 647412. https://doi.org/10.1101/647412.

Tyrmi, Jaakko, Jaana Vuosku, Juan Acosta, Zhen Li, Lieven Sterck, Maria Cervera, Outi Savolainen, and Tanja Pyhäjärvi. 2020. 'Supplemental Material for Tyrmi et al., 2020'. GSA Journals. https://doi.org/10.25387/G3.12403448.

Uchiyama, Kentaro, Tokuko Ujino-Ihara, Saneyoshi Ueno, Yuriko Taguchi, Norihiro Futamura, Kenji Shinohara, and Yoshihiko Tsumura. 2012. 'Single Nucleotide Polymorphisms in Cryptomeria Japonica: Their Discovery and Validation for Genome Mapping and Diversity Studies'. *Tree Genetics & Genomes* 8 (December). https://doi.org/10.1007/s11295-012-0508-

Vendramin G.G., B. Fady, S.C. González-Martínez, F.S. Hu, I. Scotti, F. Sebastiani, Á. Soto, R.J. Petit (2008) - Genetically depauperate but widespread: the case of an emblematic Mediterranean pine. Evolution 62: 680-688.

Vendramin Giovanni G., Bruno Fady, Hadrien Lalagüe, Milvia Racchi, Federico Sebastiani, Remy Petit, and Santiago C. Gonzalez-Martinez. 2010. 'Nuclear Microsatellites and SNPs Reveal Low Diversity in Stone Pine (Pinus Pinea L.), a Genetically Depauperated Species'. In Biodiversity International. http://agris.fao.org/agris-search/search.do?recordID=FR2014017347.

Vuosku, Jaana, Katja Karppinen, Riina Muilu-Mäkelä, Tomonobu Kusano, G H M Sagor, Komlan Avia, Emmi Alakärppä, et al. 2018. 'Scots Pine Aminopropyltransferases Shed New Light on Evolution of the Polyamine Biosynthesis Pathway in Seed Plants'. *Annals of Botany* 121 (6): 1243–56. https://doi.org/10.1093/aob/mcy012.

Vuosku, Jaana, Riina Muilu-Mäkelä, Komlan Avia, Marko Suokas, Johanna Kestilä, Esa Läärä, Hely Häggman, Outi Savolainen, and Tytti Sarjala. 2019. 'Thermospermine Synthase (ACL5) and Diamine Oxidase (DAO) Expression Is Needed for Zygotic Embryogenesis and Vascular Development in Scots Pine'. *Frontiers in Plant Science* 10. https://doi.org/10.3389/fpls.2019.01600.

Wachowiak, Witold, Urmi Trivedi, Annika Perry, and Stephen Cavers. 2015. 'Comparative Transcriptomics of a Complex of Four European Pine Species'. *BMC Genomics* 16 (1): 234. https://doi.org/10.1186/s12864-015-1401-z.

Wong, Melissa M. L., Charles H. Cannon, and Ratnam Wickneswari. 2012. 'Development of High-Throughput SNP-Based Genotyping in Acacia Auriculiformis x A. Mangium Hybrids Using Short-Read Transcriptome Data'. *BMC Genomics* 13 (December): 726. https://doi.org/10.1186/1471-2164-13-726.

Zimin, Aleksey V., Kristian A. Stevens, Marc W. Crepeau, Daniela Puiu, Jill L. Wegrzyn, James A. Yorke, Charles H. Langley, David B. Neale, and Steven L. Salzberg. 2017. 'An Improved Assembly of the Loblolly Pine Mega-Genome Using Long-Read Single-Molecule Sequencing'. *GigaScience* 6 (1): 1–4. https://doi.org/10.1093/gigascience/giw016.