

Research Paper

Scanning RNA-Seq and RAD-Seq approach to develop SNP markers closely linked to *MALE STERILITY 1 (MSI)* in *Cryptomeria japonica* D. Don

Saneyoshi Ueno^{*1)}, Kentaro Uchiyama¹⁾, Yoshinari Moriguchi²⁾, Tokuko Ujino-Ihara¹⁾, Asako Matsumoto¹⁾, Fu-Jin Wei¹⁾, Maki Saito³⁾, Yumi Higuchi⁴⁾, Norihiro Futamura¹⁾, Hiroyuki Kanamori⁵⁾, Yuichi Katayose⁵⁾ and Yoshihiko Tsumura^{1,6)}

¹⁾ Department of Forest Molecular Genetics and Biotechnology, Forestry and Forest Products Research Institute, Forest Research and Management Organization, 1 Matsunosato, Tsukuba, Ibaraki 305-8687, Japan

²⁾ Graduate School of Science and Technology, Niigata University, 8050, Ikarashi 2-no-cho, Nishi-ku, Niigata 950-2181, Japan

³⁾ Toyama Prefectural Agricultural Forestry and Fisheries Research Center, Forestry Research Institute, Yoshimine 3, Tateyama-cho, Nakashinkawagun, Toyama 930-1362, Japan

⁴⁾ Niigata Prefectural Forest Research Institute, 2249-5 Unotoro, Murakami, Niigata 958-0264, Japan (retired)

⁵⁾ National Institute of Agrobiological Sciences, Owashi, Tsukuba, Ibaraki 305-8634, Japan

⁶⁾ Faculty of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan

Cryptomeria japonica is a major forestry tree species in Japan. Male sterility of the species is caused by a recessive gene, which shows dysfunction of pollen development and results in no dispersed pollen. Because the pollen of *C. japonica* induces pollinosis, breeding of pollen-free *C. japonica* is desired. In this study, single nucleotide polymorphism (SNP) markers located at 1.78 and 0.58 cM to a male sterility locus (*MSI*) were identified from an analysis of RNA-Seq and RAD-Seq, respectively. SNPs closely linked to *MSI* were first scanned by a method similar to MutMap, where a type of index was calculated to measure the strength of the linkage between a marker sequence and *MSI*. Linkage analysis of selected SNP markers confirmed a higher efficiency of the current method to construct a partial map around *MSI*. Allele-specific PCR primer pair for the most closely linked SNP with *MSI* was developed as a codominant marker, and visualization of the PCR products on an agarose gel enabled rapid screening of male sterile *C. japonica*. The allele-specific primers developed in this study would be useful for establishing the selection of male sterile *C. japonica*.

Key Words: conifer, gymnosperm, non-model organisms, pinales, expressed sequences, restriction-site-associated DNA, marker-assisted selection.

Introduction

Male sterility in sugi (*Cryptomeria japonica*) was first described by Taira *et al.* (Taira *et al.* 1993) in Toyama prefecture, Japan, and increased efforts to identify sterile trees led to the discovery of 23 male sterile sugi individuals from all over Japan (Saito 2010). These sterile trees have one of four sterile alleles of *MS1*, *MS2*, *MS3*, and *MS4*, which show Mendelian inheritance (Miyajima *et al.* 2010, Taira *et al.* 1999, Yoshii and Taira 2007). Because no mature pollen is produced and dispersed from sterile trees, they help decrease pollen allergy, which causes heavy social and economic losses every year in Japan. The breeding of male

sterile trees would be accelerated if marker assisted selection (MAS) was possible. To assure the precision of selection, the distance between molecular markers and a candidate gene has to be close enough.

Long-lasting efforts to develop molecular markers in sugi have resulted in a high-density linkage map with 2,560 markers in 11 linkage groups and 1,266.2 cM in length (Moriguchi *et al.* 2016). The majority of these markers are based on randomly selected SNPs from expressed sequence tags (ESTs) or transcriptome contigs (Ueno *et al.* 2012) by RNA Sequencing (RNA-Seq). However, randomly designing molecular markers and searching for those linked with target genes are inefficient ways to identify male sterile individuals with MAS (Moriguchi *et al.* 2017). To overcome such drawback, MMAPPR (Hill *et al.* 2013) and a series of MutMap methods (Abe *et al.* 2012, Fekih *et al.* 2013, Takagi *et al.* 2013) have been developed using the association between a target phenotype and candidate SNPs over

Communicated by Hiroyoshi Iwata

Received December 14, 2017. Accepted October 1, 2018.

First Published Online in J-STAGE on February 15, 2019.

*Corresponding author (e-mail: saueno@ffpri.affrc.go.jp)

the genome before developing genetic markers.

In MutMap, SNP index is calculated as the “ratio between the number of reads of a mutant SNP and the total number of reads corresponding to the SNP” (Abe *et al.* 2012) in a bulked DNA sequencing of mutant and wild-type phenotypes. When we consider a simple Mendelian recessive trait in a backcross family, the SNP indices from mutant and wild-type individuals are expected to be approximately 1.0 and 0.5, respectively. The original MutMap method utilizes whole genome sequencing reads from bulked samples and needs a reference genome sequence of a target species. Although an increasing number of genome sequences have been published in recent years, construction of a reference genome sequence for sugi, a species with large genome (11 Gb) (Hizume *et al.* 2001), and repetitive elements, has been still difficult.

There are several methods of analysis to reduce the complexity of a genome. One of these includes the use of ESTs (Adams *et al.* 1992) or a transcriptome (Velculescu *et al.* 1997), whereby only expressed parts of a genome are sampled for downstream analysis. Another method to reduce the genome complexity is a construction of a reduced representation library from genomic DNA, such as AFLP (Vos *et al.* 1995) and RAD-Seq (Baird *et al.* 2008) techniques. In these systems, restriction enzymes are used to cut specific recognition sites with four to eight bases, resulting in a collection from a small portion of a genome. These methods of complexity reduction by sampling specific parts of a genome may help genetic analysis for species with a large genome and without a reference genome sequence.

In this study, we applied a method similar to MutMap to develop SNP markers that are closely linked to a male sterility gene using transcriptomic reference as well as genomic pseudo-reference constructed from RAD-Seq for sugi. Then, we screened and identified markers closely linked to one of the male sterile genes (*MSI*) in the genome, without any reference genome sequences available to date.

Materials and Methods

Plant materials and phenotyping of male sterility

The mapping pedigree (T5 family) consisted of 173 individuals and was derived from a backcross between ‘Toyama-1’ and ‘T1NK4F1’ trees (‘T1NK4F1’ is male fertile, *Ms1/ms1*) (Moriguchi *et al.* 2014). ‘T1NK4F1’ was an F₁ hybrid between ‘Toyama-1’ (male sterile, *ms1/ms1*) and ‘Nakakubiki-4’ (male fertile, *Ms1/Ms1*). Male flowers were induced by gibberellic acid according to the manufacturer (KYOWA HAKKO BIO CO., LTD, Tokyo, Japan) instructions. Pollen production in each individual was visually assessed using a stereomicroscope in the following spring. Induction of male flowers and phenotyping of pollen production were conducted at Niigata Prefectural Forest Research Institute (Murakami, Niigata, Japan).

RNA-Seq from male flower buds

The T5 family at the age of three years was planted in a nursery at the Forestry and Forest Products Research Institute, Tsukuba, Japan, in April 2011. Male flowers were induced by applying gibberellic acid according to the manufacturer’s (KYOWA HAKKO BIO CO., LTD) instructions in July. On the 6th, 11th, 14th, 19th, and 24th of October 2011, male flower buds from sterile and fertile individuals were sampled. Due to the limited availability of male flower buds from each individual, we sampled two to three of them from each individual and mixed together with buds from other individuals. We mixed buds up to 50 individuals in each phenotype category (sterile and fertile). Male flower bud samples were stored at -80°C until RNA extraction. They were then powdered in liquid nitrogen with mortar and pestle, and total RNA was extracted by the method of Le Provost *et al.* (Le Provost *et al.* 2007). About 10 spoonfuls of the powder were added to 1 mL of CTAB extraction buffer. The extracted RNA was further purified by SV total RNA purification system (Promega, Madison WI). For sequencing, an equal amount of RNA from each sampling date was mixed for each category (sterile and fertile), and the sequencing library was constructed by Hokkaido System Science Co., Ltd. (Sapporo, Japan) and then sequenced on a MiSeq (Illumina, San Diego, CA) in 2×250 bp paired ends at the National Agriculture and Food Research Organization, Tsukuba, Japan. Average depth was insufficient in a preliminary read mapping to full length cDNA (FLcDNA) sequences (Futamura *et al.* 2008) (**Supplemental Text 1**). Only 9,478 (42.0%) FLcDNA sequences had more than $20\times$ average depth. Therefore, additional sequence was collected for each category (sterile and fertile) with HiSeq2000 (Illumina) in 2×101 bp paired-end sequencing by Hokkaido System Science Co., Ltd.

RAD-Seq

DNA was extracted from each individual of the T5 family according to a modified CTAB method (Doyle and Doyle 1987). A total 20 individuals (8 sterile and 12 fertile trees) and the parents (‘Toyama-1’ and ‘T1NK4F1’) were used in the double digest RAD-Seq (ddRAD) experiment (Peterson *et al.* 2012). Briefly, genomic DNA was double-digested by restriction enzymes, *SphI* and *MluCI*. The digested fragments were size-selected using BluePippin agarose gel (Sage Science, Beverly, MA) and ligated with adapters (Fluidigm, South San Francisco, CA). After PCR amplification with adapter-specific primer pairs, an equal amount of DNA from each sample was mixed to construct a RAD library. The quality of the library was checked by sequencing on MiSeq in 2×300 bp paired ends and finally sequenced in 2×101 bp paired ends on HiSeq2000 by Hokkaido System Science Co., Ltd.

Selection of candidate markers

RNA-Seq reads were cleaned with CLC Genomics Workbench (GW) (Qiagen, Hilden, Germany) 7.0.3 or

Trim_Galore program (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The cleaned reads were mapped against 22,539 (36.0 Mb) full-length cDNA (FLcDNA) sequences (Futamura *et al.* 2008) (**Supplemental Text 1**) using GW 8.5.1 with more stringent parameters than those of default (length fraction = 0.85 and similarity fraction = 0.9). Read mappings were carried out for each sample and merged into a BAM alignment file for sterile and fertile samples. Candidate SNP sites were detected from the BAM files using VarScan (ver. 2.4.3) (Koboldt *et al.* 2012) with default parameter settings (min coverage = 8, min reads = 2, min varfreq = 0.2, min avgqual = 15, and P value threshold = 0.01) to generate a VCF (variant call format) file. Read depths (number of reads) for both alleles (X and Y) at any SNP sites were extracted from the VCF file.

SNPs from RNA-Seq were also searched from contigs assembled *de novo* by Trinity (Grabherr *et al.* 2011) using the cleaned RNA-Seq reads. The contig sequences were used as references, onto which the reads were mapped, and the SNPs were detected as described above using the same parameter settings.

For RAD-Seq, reads from each individual were combined according to male flower phenotypes (sterile and fertile), which were then processed by dDocent (version 2.17) pipeline (Puritz *et al.* 2014). In the reference construction in dDocent, we set “CUTOFF” of four reads within a sample, and “CUTOFF2” of two which indicated that at least two samples (sterile and fertile samples in the present case) had that contig sequence. After mapping the reads onto the references, the resulting BAM files were then used to detect SNPs, and the read depth at each SNP site was extracted in the same manner as done for the abovementioned RNA-Seq data.

To select candidate markers linked to *MSI*, we applied a procedure similar to MutMap. The read depth for each allele at each SNP site was assumed to be proportional to the number of alleles in the genotype. Heterozygous (XY) and homozygous (XX) genotypes were expected to produce reads from X allele in ratios of 50% and 100%, respectively, to total reads mapped at each focal SNP site. The difference (*D*) in the X allele percentage between heterozygous (fertile) and homozygous (sterile) samples was calculated for each SNP site and averaged over all SNP sites on the same reference sequence (full length cDNA (FLcDNA), RNA-Seq, or RAD-Seq reference). The average value is referred to as *Dav* hereafter. Only the *Dav* was calculated for the reference with at least two SNP sites and with a maximum distance among all SNPs of ≥ 50 bp (**Supplemental Fig. 1**). The threshold distance among SNPs was determined arbitrarily as 50 bp in the current study. Because *C. japonica* is allogamous, we expected four possible cases for linkage phase of a two-alleles SNP site and *MSI* and the resulting values of *Dav* in our backcross family (**Fig. 1**). Theoretically, a maximum of three alleles in a SNP site is possible in backcross family, but these cases are expected to be rare, and we focused on cases for two-alleles SNP sites here. The

frequency for three-alleles SNP sites was estimated as 0.66% with four screening panel individuals in our previous re-sequencing study (Uchiyama *et al.* 2012). In the current procedure described above, we focused on cases where *Dav* ≈ 0.5 (**Fig. 1B, 1D**). Simple simulation, where a Mendelian marker is in complete linkage equilibrium with *MSI*, indicated that the average difference ($\bar{D}$) was less than 0.1, when the number of backcross samples exceeded 20 (*msl/msl:MSl/MSl* = 10:10) (**Supplemental Fig. 2, Supplemental Text 2**). In the current study, about 50 male sterile and 50 male fertile samples were included in the RNA-Seq experiment, whereas the RAD-Seq had nine sterile and 13 fertile samples including their parents.

SNP assay and linkage mapping

The SNPs on some of the reference sequences with higher *Dav* values were assessed by the EP1 system (Fluidigm) (Wang *et al.* 2009). We selected SNPs with higher *Dav* values sequentially, but if there were any adjacent SNPs within 30 bp of target SNP, it was not suitable for SNPType Assay Design (Fluidigm) and discarded. In total, 48, 33, and 48 SNP assays from FLcDNA, RNA-Seq, and RAD-Seq reference sequences were tested. We used T5 family (173 individuals) for linkage mapping. DNA was extracted for these individuals from the needle tissue using a modified CTAB method (Doyle and Doyle 1987) and used for SNPType assay by the Fluidigm EP1 system according to the manufacturer's instructions. Fluidigm SNP Genotyping Analysis (version 4.1.2) was used for genotype calling under the default setting.

For linkage mapping, SNP genotypes were combined with 27 marker genotypes we had obtained in our previous study (Moriguchi *et al.* 2016). After removing loci with a large number of missing data points (more than 10 individuals) or extreme segregation distortion ($P < 0.001$), a partial linkage map for the ninth linkage group (LG9) where the male sterility gene (*MSI*) exists, was constructed by JoinMap v4.0 (Kyzama, Wageningen, The Netherlands) with the parameter BC (backcross) and two rounds of map calculation according to Moriguchi *et al.* (2016). Map distance was calculated by the Kosambi mapping function, and the linkage map for LG9 was drawn by MapChart ver. 2.0 (Voorrips 2002).

Allele-specific PCR and agarose gel electrophoresis

After identification of the most closely linked markers to *MSI* on the partial linkage map, we constructed allele-specific PCR (ASP) system or Amplification Refractory Mutation System (ARMS) (Newton *et al.* 1989) for an easy identification of male sterile trees by agarose gel electrophoresis. The ASP primers were designed using Primer3Plus (Untergasser *et al.* 2007) based on the method described by Wangkumhang *et al.* (2007). Template sequence for primer design contained two SNPs, and we designed ASP primers for both of these (**Fig. 2**). Because two SNPs were located on 3' and 5' ends of the sequence, it was not possible to

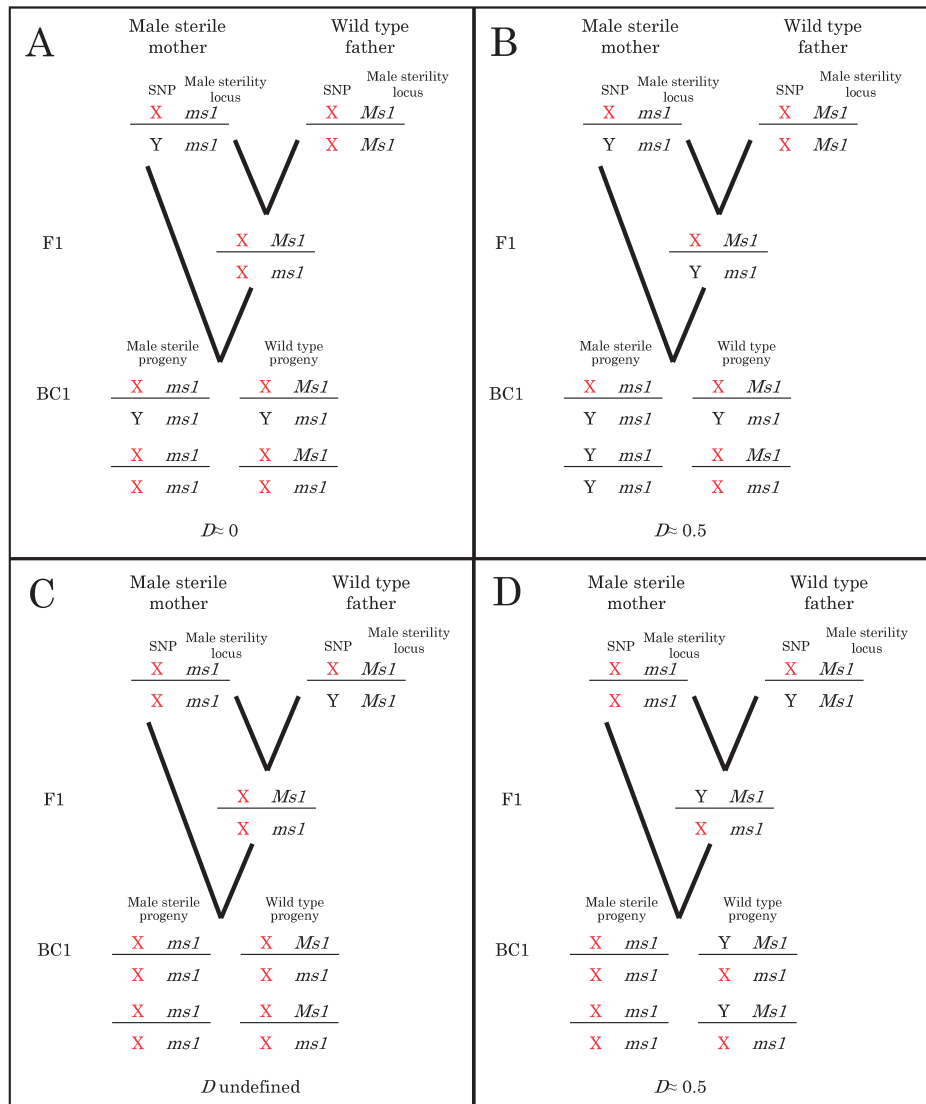


Fig. 1. Schematic representation of calculation of D at a SNP site with two alleles and linkage phase between male sterility locus (MSI) in a backcross family. “X” and “Y” represent SNP alleles. No recombination is assumed between the SNP site and MSI in this example. D_{av} is an averaged value of D over SNP sites in a reference sequence (contig). D is defined as the difference in X allele frequency between male sterile and wild-type pedigree members. Four possible cases (A–D) in terms of allelic linkage between an SNP site and MSI in backcross scheme are presented. In case C, the SNP alleles in the parents are not separated in the mapping family and D cannot be calculated.

design ARMS primers (Newton *et al.* 1989), which target both alleles in a SNP with oppositely directed allele-specific primers. The reaction mixture, totaling 10 μ L, contained 3 μ L of 2 $\times$ Multiplex (Qiagen), 0.2 μ M of each forward and reverse primers, and 1 μ L (5–10 ng) of DNA, which was then amplified in a GeneAmp PCR System 9700 (Applied Biosystems, Foster City, CA) with the following thermal conditions: initial denaturation at 95°C for 15 min, 4 cycles of 95°C for 30 s, 64°C for 90 s with -1° C per cycle and 72°C for 30 s, and 34 cycles of 95°C for 30 s, 60°C for 90 s, and 72°C for 30 s. The PCR products were visualized by 2% agarose gel electrophoresis and ethidium bromide staining.

Registration of sequence data

All the sequence data used in the present study were registered in DDBJ with the accession numbers mentioned in **Supplemental Text 1**.

Results

Scanning RNA-Seq

In total, 49.6 Gb of transcriptome reads were collected (5.0 Gb from MiSeq and 44.6 Gb from HiSeq). As a preliminary study, MiSeq collected 2.7 Gb from male sterile and 2.3 Gb from male fertile flower bud samples. Due to adapter or low-quality sequences, 40.1% of bases were removed. Initial mapping of these reads onto FLcDNA reference

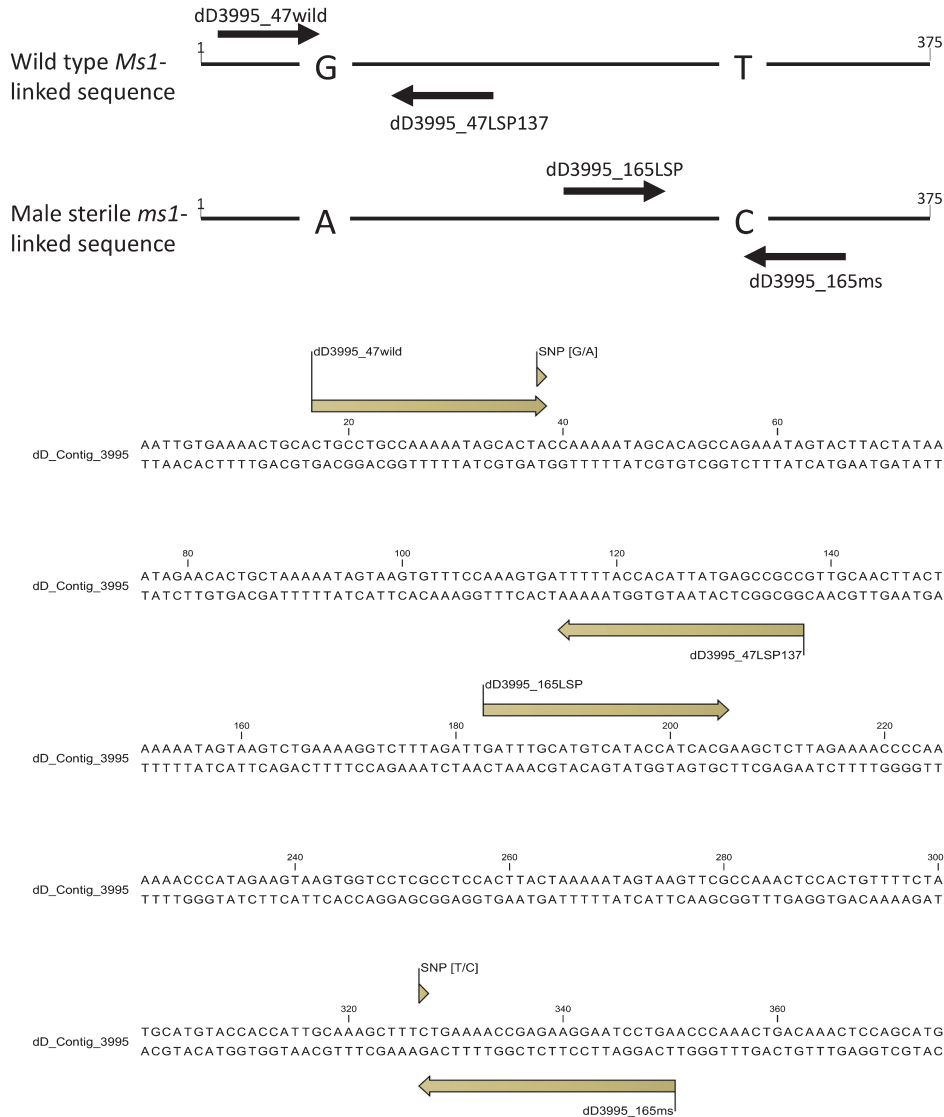


Fig. 2. Allele-specific PCR primers to discriminate male sterile and fertile individuals with agarose gel electrophoresis. There are two SNP sites on dD_Config_3995 measuring 375 bp in length. Each SNP site is used to target a specific SNP allele coupled with either *Ms1* (wild-type) or *ms1* (male sterile).

sequences showed that 67.0% of the reads were successfully mapped, with $32.3\times$ average depth. Only 9,478 (42.0%) FLcDNA sequences had more than $20\times$ average depth. We therefore collected additional reads using HiSeq2000, which produced 22.9 and 23.2 Gb sequencing reads from fertile and sterile samples, respectively. After the removal of 1.41 Gb (3.1%) of reads by Trim_Galore program, the remaining 22.1 and 22.5 Gb reads from sterile and fertile samples, respectively, were mapped against FLcDNA reference, resulting in 30.8 Gb (68.2%) of mapped bases with $507.6\times$ average depth. In addition, 21,457 (95.2%) FLcDNA sequences reached more than $20\times$ average depth. We identified 223,726 candidate variant (223,418 SNPs and 13,369 indels) sites in total, which indicated 9.9 variations per reference, and the ratio of transition- and transversion-type SNPs (Ts/Tv) was 1.49. A total 14,072 (62.4%) references

had two or more SNPs, and the maximum distance among the SNPs was ≥ 50 bp, for which *Dav* was evaluated, ranging from 0.0 to 0.95 with an average of 0.112.

De novo assembly of the cleaned HiSeq reads produced 232,103 contigs with a total length of 177.5 Mbp, onto which cleaned reads from sterile and fertile samples (22.1 and 22.5 Gb, respectively) were separately mapped, resulting in 42.9 Gb (96.1%) of the mapped bases with a $126.7\times$ average depth. In total, 46,303 (19.9%) RNA-Seq contigs had more than $20\times$ average depth. We identified 517,225 candidate variant (517,140 SNPs and 85 indels) sites, indicating 2.2 variations per reference (i.e. 517,225 candidate variants per 232,103 contigs), and the ratio of transition- and transversion-type SNPs (Ts/Tv) was 1.59. A total of 40,573 (17.5%) references had two or more SNPs, and the maximum distance among the SNPs was ≥ 50 bp, for which

Dav was evaluated, ranging from 0.0 to 0.95 with an average of 0.121.

Scanning RAD-Seq

We collected 2.3 and 2.8 Gb of sequences from sterile and fertile trees, respectively. The dDocent pipeline removed 712 Mb (13.8%) of adapter or low-quality bases, producing 16,455 reference contigs (3.3 Mb in length). Most of reference sequences were 212 bp in length with 10 “N”s in the middle, where forward and reverse reads were joined, but 992 of the references were merged contigs from forward and reverse reads. The pipeline then mapped 979.9 Mb (22.0%) of reads onto the reference. In total, 98,887 candidate SNP sites were detected from 9,183 (55.8%) contigs, resulting in an average number of 10.8 SNPs per reference, and the ratio of transition- and transversion-type SNPs (Ts/Tv) was 2.56. A total of 6,638 references had two or more SNPs, and the maximum distance among the SNPs was ≥ 50 bp, for which *Dav* was evaluated, ranging from 0.0 to 0.70 with an average of 0.113. The depth for these SNP sites ranged from 5 to 4,013 with $85.9\times$ average depth.

SNP assay and linkage mapping

Three sets of SNPTYPE assays (129 SNPs including 48, 33 and 48 SNPs from FLcDNA, RNA-Seq, and RAD-Seq references, respectively) were constructed based on *Dav*, whose ranges were 0.013–0.54, 0.37–0.50, and 0.30–0.56, respectively for FLcDNA, RNA-Seq and RAD-Seq references (data not shown). The assay was successful [number of missing genotypes ≤ 10 , and no severe segregation distortion ($P \geq 0.001$)] for 29 (60.4%) 24 (72.7%), and 22 (45.8%) SNPs originating from FLcDNA, RNA-Seq, and RAD-Seq, respectively. Linkage analysis using JoinMap in-

cluded 26 backcross-type markers (eight, 13, and five from FLcDNA, RNA-Seq and RAD-Seq, respectively) and indicated that 16 markers were mapped onto LG9, where *MSI* exists (Table 1, Fig. 3). The closest markers were CSFL022_N13-194, c65725_g1_i1-123, and dDContig_3995-165 from FLcDNA, RNA-Seq, and RAD-Seq, showing 2.95, 1.78, and 0.58 cM to *MSI*, respectively. The total LG9 map distance increased by 6.8 cM with the 16 markers added to the previous map (Moriguchi *et al.* 2016) (Supplemental Fig. 3).

Allele-specific PCR system using agarose gels

Two pairs of allele-specific PCR primers (Table 2, Fig. 2) were designed targeting dD_Contig_3995, which was the most closely linked to *MSI* based on the abovementioned

Table 1. SNP markers on linkage group 9 (LG9) around male sterility locus (*MSI*) in *C. japonica*

Marker origin	Marker ID	<i>Dav</i>	Map position on LG9 (cM)	Distance to <i>MSI</i> (cM)
FLcDNA	CLFL025_C07-615	0.346	45.44	23.52
	CFFL008_J05-175	0.231	44.28	22.35
	CFFL025_E16-1334	0.422	5.90	16.03
	CLFL044_G04-1079	0.247	5.90	16.03
	CMFL035_K17-90	0.334	13.71	8.22
	CLFL018_I12-1304	0.515	17.24	4.69
	CSFL023_N16-2950	0.412	17.24	4.69
	CSFL022_N13-194	0.454	18.99	2.94
RNA-Seq	c76407_g8_i1-82	0.371	10.12	11.81
	c76501_g12_i1-230	0.445	10.12	11.81
	c60295_g2_i1-1291	0.480	17.24	4.69
	c66322_g1_i7-1924	0.483	18.99	2.94
	c65725_g1_i1-123	0.432	20.16	1.77
RAD-Seq	dDContig_13452-169	0.329	70.36	48.44
	dDContig_6119-92	0.422	28.01	6.09
	dDContig_3995-165	0.467	22.51	0.58
Fertility	<i>MSI</i>	NA	21.93	0.00

Dav represents levels of linkage between marker and *MSI*. The *Dav* value of 0.5 is a theoretical prerequisite for complete linkage with *MSI*. NA: Not available

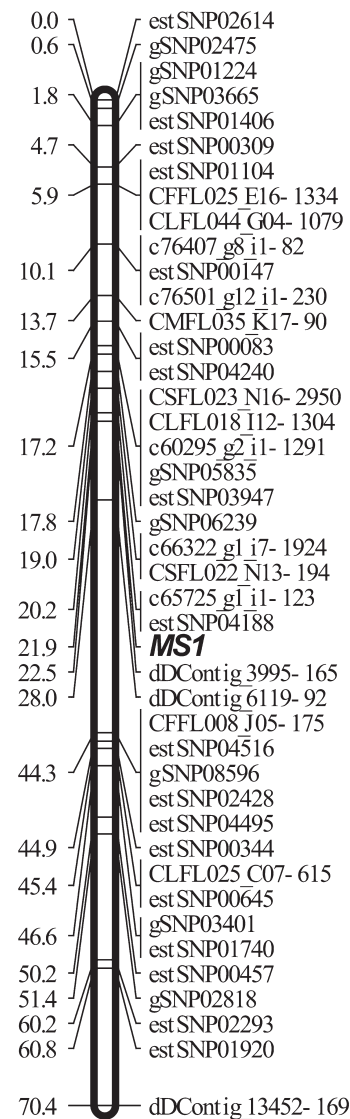


Fig. 3. A partial linkage map for linkage group 9 (LG9) around the male sterility locus *MSI*. Marker names and centimorgan distances (Kosambi) are indicated to the right and left of the linkage group, respectively.

Table 2. Allele-specific PCR primers for dD_Contig_3995 discriminating male sterile and fertile trees on agarose gels (Figs. 2, 4)

Primer ID	Primer sequence [#]	Expected PCR product size (bp)	Coupling allele
dD3995_47wild	CTGCCTGCCAAAATAGCACGG	121	<i>MsI</i>
dD3995_47LSP137	GGCGGCTCATAATGTGGTAAAAA		
dD3995_165ms	TCAGGATTCTTCTCGGTTTTCCG	168	<i>msI</i>
dD3995_165LSP	GATTTCATGTCATACCATCACG		

[#] Underlined bases indicate artificial mismatch introduced to increase specificity.

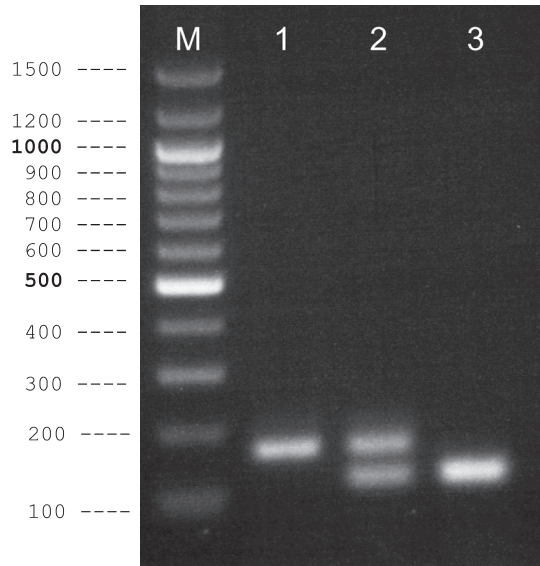


Fig. 4. Agarose gel electrophoresis to discriminate male sterile and fertile *C. japonica* in T5 family. Sample DNAs were amplified using primers shown in Table 2. PCR products were separated on 2% agarose gel and visualized by ethidium bromide staining. M: DNA Ladder One (nacalai tesque, Kyoto, Japan) (numbers on the left are molecular sizes in base pairs), 1: ‘Toyama-1’ with *msI/msI*, 2: ‘T5NK4F1’ with *MsI/msI*, 3: ‘Nakakubiki-4’ with *MsI/MsI*.

mapping result. First, we used a single allele-specific primer pair in a single PCR reaction and confirmed allele-specific amplification from *MsI/MsI*, *MsI/msI*, and *msI/msI* individuals (data not shown). Two pairs of allele-specific PCR primers were then combined, and multiplex PCR was performed, which enabled the detection of both male sterile- and wild-type alleles in a single PCR reaction (Fig. 4).

Discussion

Detection and validation of candidate SNPs

SNP detection is a fundamental step to develop genetic markers. We detected 223,726, 517,140, and 98,887 candidate SNPs from the FLcDNA, RNA-Seq and RAD-Seq references, respectively. A total of 129 were selected for validation to perform a Fluidigm SNPTyping assay, which resulted in 75 (58.1%) successful markers. When the genotyping scatter plots for unsuccessful markers were inspected, some of the markers appeared to display a crushed or monomorphic pattern (Supplemental Fig. 4). Markers with

a monomorphic pattern (seven, one and one candidate SNP from FLcDNA, RNA-Seq, and the RAD-Seq reference, respectively) indicates that we detected false SNP sites, which may have occurred due to a sequencing or data analysis error, null alleles (amplification failure in PCR due to SNPs at the primer binding site or third SNP allele), duplicated genomic sequences (paralogues), or RNA editing (Piskol *et al.* 2013). The higher monomorphic pattern rate in the candidate SNPs from the FLcDNA reference suggests that some of the reads were forced to be mapped onto paralogous sequences.

In the coding region of *C. japonica*, one exon SNP for every 106.7 bp was reported by re-sequencing four genotypes with the Sanger method (Uchiyama *et al.* 2012). The candidate SNP frequency from 35.96 Mbp of the FLcDNA reference in the current study based on 223,418 SNPs was roughly estimated as one candidate for every 161 bp. Although we detected 517,140 candidate SNPs from 177.5 Mb of RNA-Seq reference, most (80.1%) of them had less than 20× average depth, for which the SNP frequency calculation may be inappropriate. The ratio of transition- and transversion-type SNPs (Ts/Tv) can be considered an SNP quality metric because if all of the mutations occur at the same probability, we expect a Ts/Tv ratio ≈0.5, while the transversion type mutation rate is lower in general and Ts/Tv ratios of 2.15 and 3.27 has been reported in the final call set for human genome-wide and exome known SNPs, respectively (DePristo *et al.* 2011). Similar Ts/Tv ratio values (1.21–1.51) as the current study (1.49 and 1.59) have been reported in RNA-Seq data from coniferous species (Chen *et al.* 2012, Plomion *et al.* 2016, Ueno *et al.* 2018).

To gain further insights into the validity of the candidate SNPs, the RAD-Seq reads were mapped onto the FLcDNA references (hereinafter referred to as “RAD-FL mapping”). SNPs were detected by VarScan (ver. 2.4.3) (Koboldt *et al.* 2012) with default parameter settings, and the resulting VCF file was compared with that created from RNA-Seq reads’ mapping onto the same FLcDNA reference (hereinafter referred to as “RNA-FL mapping”). The comparison was made using “bcftools stats” and plot-vcfstats in bcftools ver. 1.7-1 (Li and Durbin 2009). In total, 2,014 candidate SNPs were common between RAD-FL and RNA-FL mapping. These candidates were distributed among 456 (2.0%) FLcDNA reference sequences. The Ts/Tv ratio for these common SNPs was 2.03. The mapping of RAD-Seq reads onto a genomic reference helps validate the SNPs detected by RNA-Seq.

Scanning linked markers from pooled data

Male sterility in *C. japonica* has been investigated with an aim of identifying the responsible genes of the phenotype and for a better understanding of the molecular mechanisms of male sterility (Futamura *et al.* 2008, Moriguchi *et al.* 2012, 2016, Tsubomura *et al.* 2016). Analysis of EST (Expressed Sequence Tag) or RNA-Seq with high-throughput sequencers are both a method to capture the whole atlas of expressed genes (transcriptome) in a tissue (Weber 2015). Differentially expressed genes between male fertile and sterile strobili have been identified (Futamura *et al.* 2011, Tsubomura *et al.* 2016). Although the male sterility in *C. japonica* is caused by a single gene, none of the differentially expressed genes has directly dictated the phenotype and linkage analysis indicates that they are different from *MSI* (Ihara *et al.* unpublished results). In the current study, we focused on the SNPs detected from high-throughput sequencing in the mapping family of *MSI* and identified closely linked markers (Fig. 3).

Due to the backcross family we used, fertile and sterile genotypes are heterozygotes and homozygotes, respectively. If we had used other families such as selfed mapping family, homozygous fertile and sterile individuals would have been separated. Homozygous genotypes are easier to genotype, and larger values of expected score ($Dav \approx 1$) would be preferable to differentiate both phenotypes. Unfortunately, in most of conifer tree species, which are allogamous, selfing leads to serious inbreeding depression, and construction of such mapping families is difficult. Heterozygote genotype calling is therefore required in allogamous species, which needs a higher (15×, for example) average depth (Song *et al.* 2016).

Because of the expensive sequencing costs to perform individual genotyping, especially for RNA-Seq, pooling strategy helped to reduce the cost, and selecting linked SNPs with appropriate Dav increased the efficiency of developing the markers. Of 21 analyzed SNPs from FLcDNA and RNA-Seq in the current study, 13 (62%) SNPs were mapped onto LG9 (Table 1). If we had randomly selected SNPs, a simple calculation would suggest 9.1% of markers on the chromosome, because *C. japonica* has 11 chromosomes. Considering the size of LG9 (96.9 cM or 7.7% of total map distance) (Moriguchi *et al.* 2016), the number of markers expected on LG9 may be <9.1%. Scanning RNA-Seq data increased the efficiency of partial linkage map construction by more than six times, compared with the method of developing markers randomly from the entire genome. Similarly, RAD-Seq data resulted in three (60%) linked markers of five analyzed SNPs. In the current study, we designed SNP assays with Dav ranges of 0.013–0.54, 0.37–0.50, and 0.30–0.56 for FLcDNA, RNA-Seq and RAD-Seq references, respectively, to design 129 Fluidigm SNPtype assay primer pairs in total. If we had selected assays with $Dav > 0.4$, for example, the efficiency to construct the partial linkage map would have been higher.

To gain some insight into Dav critical threshold values,

the values were calculated for 771 (Moriguchi *et al.* 2012) and 2,160 (Moriguchi *et al.* 2016) expressed genic markers over 11 linkage groups, which were mapped in the YI mapping family (Supplemental Text 3). As expected, markers in LG9 where *MSI* existed revealed higher Dav values, although *MSI* was not mapped on the YI map and its exact location was unclear. If we defined the Dav critical threshold values as 1% of the upper tail of the Dav distribution, they were 0.343 and 0.312 for markers from Moriguchi *et al.* (2012) and Moriguchi *et al.* (2016), respectively, and the rate of false positives was <50%. These results agree well with the observed results (Table 1), where 7 (33.3%) of 21 assayable markers from the FLcDNA and RNA-Seq references were not mapped on LG9. Notably, the empirical estimates of the Dav distribution and the rate of false positives in Supplemental Text 3 were based on RNA-Seq reads. If we apply scanning RAD-Seq for the empirical estimates, these critical threshold values may change. The broader the Dav range considered, the larger the number of markers linked to *MSI*. However, in practice, the Dav threshold values are determined by the cost and time required for the specific assay method.

In the preliminary simulation (Supplemental Fig. 2), the difference (D) (for Fig. 1D) was calculated in cases where there was no linkage between a marker and *MSI* in the backcross (BC1) pedigree, indicating that D approached zero as the number of BC1 offspring (N) increased, and that the average difference ($\bar{D}$) was <0.1 when $N > 20$. Overall concordance was observed when the observed D values were compared with the simulated values. The observed/simulated $\bar{D}$ values were 0.1050/0.0456, 0.0995/0.0456, and 0.1035/0.0987 for SNPs from FLcDNA ($N = 100$), RNA-Seq ($N = 100$), and RAD-Seq ($N = 20$), respectively (Supplemental Fig. 5). The $\bar{D}$ values for the observed values may be higher than those for the simulated values, because some of the linked markers were included in the real situation. The observed D values from RAD-Seq are closer to the simulated values, while those based on gene expression (FLcDNA and RNA-Seq) are not. This may have been caused by unequal contribution from each allele during gene expression and/or a contribution from paralogs.

In the present study, Dav values with at least two SNPs and a maximum distance among all SNPs ≥ 50 bp on a reference were considered (Supplemental Fig. 1), because we were afraid that $Dav \approx 0.5$ may be obtained by chance even though there was no linkage between the markers and *MSI*. However, as the simple simulations (Supplemental Fig. 2) indicated that the D values (which could be considered as Dav for a reference sequence with a single SNP site) rarely exceeded 0.4 when at least 10 individuals were pooled from each category. If we considered all Dav values including Dav from a sequence with a single SNP site, the number of Dav , from the values of which linkage was evaluated, would have increased by 3,778 (23.6 %), 14,756 (36.4 %), 2,456 (37.0 %) for cases from FLcDNA, RNA-Seq, and RAD-Seq, respectively. To take advantage of many collected

reads as possible and to avoid overlooking linked markers to target character (*MSI*), *Dav* for a reference sequence with a single SNP site should be considered (**Supplemental Text 3**).

The larger the *Dav* is, the closer the marker was to *MSI* (**Table 1**). Regression of the distance (cM) to *MSI* on the *Dav* value was significant ($P = 0.0119$, **Supplemental Fig. 6**). Markers in which *Dav* > 0.475 had a high probability to link to *MSI* within 5 cM. Mishima *et al.* (2018) reported the SNP marker (reCj19250_1927) for the putative causative *MSI* gene, which showed homology to AT5G62190.1 (PRH75 DEAD box RNA helicase). The *Dav* value for reCj19250 (which corresponds to the RNA-Seq contig reference of c66093_g1_i1 in the current study) was 0.517. These data support the validity of the scanning method used in the present study. This method may be valid to search closely linked markers not only to qualitative trait loci, but also to quantitative trait loci with the largest effect.

Although pooling the sequence reads according to the phenotype to screen linked SNP candidates reduced the cost of sequencing in this study, the development of new sequencing technologies could lead to a continuous reduction in the sequencing cost in an unprecedented manner (Delseny *et al.* 2010) and make individual genotyping by sequencing (GBS) a routine task in breeding research. However, variation in the amount of sequencing reads for each individual causes uncertain genotypes for samples with low data volume, which is still problematic. Similarly, RNA-Seq data include expressional variation, and genes or alleles with low expression also cause biased data. In our preliminary analysis, we attempted to directly call the individual genotypes from RAD-Seq, estimated the error rate assuming Mendelian inheritance (Utsunomiya *et al.* 2013), and found an average Mendelian error (allelic drop in) rate of 6.9% (data not shown). Genotyping error inflates map distance and may lead to an incorrect marker order along chromosomes. Developing closely linked markers needs accurate linkage maps, and map construction from genotypes from GBS with a high error rate is probably inappropriate. Instead, scanning closely linked SNPs from pooled sequencing reads, at first, is robust against errors, and the SNP-specific markers developed achieve higher precision and scalability (from one sample to thousands of samples), enabling marker-assisted selection in the downstream applications.

Utility of the linked marker to *MSI*

The most closely linked marker developed in the present study (dD_Contig3995-165) was located 0.58 cM to the male sterility locus *MSI* (**Fig. 3**). The results were obtained from the genotypes of the mapping family. One of the 173 pedigree members showed recombination between *MSI* and the marker. This implies that discrimination of male sterile and fertile individuals based on the marker from 1,000 seedlings produced by backcross contains five to six individuals of false identification ($1,000/173 \approx 5.8$). More information with addition of a second marker located on the opposite

side of *MSI* to dD_Contig3995-165 will increase the accuracy of phenotype prediction. Selection of male sterile individuals at the embryo stage is required for mass production of male sterile trees by tissue culture from immature zygotic embryos, where marker-assisted selection (MAS) is a prerequisite. The marker developed in this study (**Table 2**) is useful in such a project, which is now in progress. The discovery of two closely linked markers that cross the target gene (*MSI*) increases the precision of MAS, while the method developed in this study will be useful for identifying other male sterile genes and their closely linked markers.

Acknowledgments

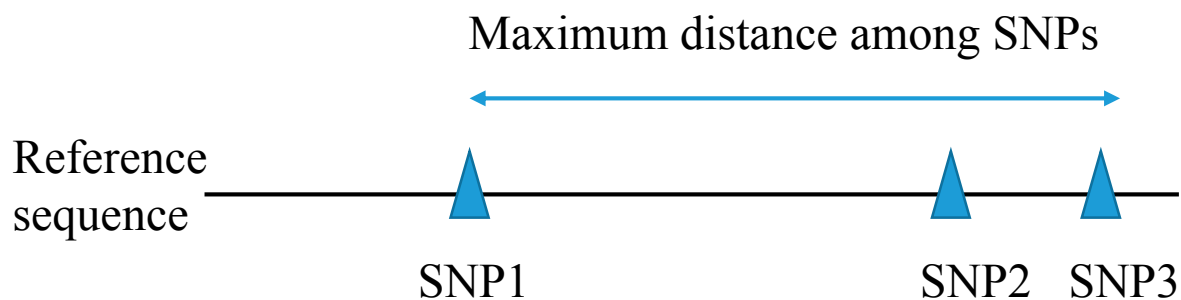
The authors would like to thank Misako Koshiba, Yasuyuki Komatsu, Nozomi Ohmiya, and Yuriko Taguchi for their assistance with laboratory work and the sampling of male flower buds in the nursery. Authors also thank two anonymous reviewers and the managing editor to improve earlier draft of the manuscript. The mapping family used in the current study was produced by Shinji Ito. Arboretum and Nursery Office, FFPRI was involved in the maintenance of *C. japonica* nursery. This work was partly supported by a Grant-in-Aid from the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry, JSPS KAKENHI (Grant Numbers 25850116 and 25450223), Research grants #201119, #201421, and #201414 from the Forestry and Forest Products Research Institute, genome-support grant from the National Institute of Agrobiological Sciences (NIAS), and grants from the Project of the NARO Bio-oriented Technology Research Advancement Institution (Research program on development of innovative technology).

Literature Cited

- Abe, A., S. Kosugi, K. Yoshida, S. Natsume, H. Takagi, H. Kanzaki, H. Matsumura, C. Mitsuoka, M. Tamiru, H. Innan *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30: 174–178.
- Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields and J.C. Venter (1992) Sequence identification of 2,375 human brain genes. *Nature* 355: 632–634.
- Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, W.A. Cresko and E.A. Johnson (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3: e3376.
- Chen, J., S. Uebbing, N. Gyllenstrand, U. Lagercrantz, M. Lascoux and T. Kallman (2012) Sequencing of the needle transcriptome from Norway spruce (*Picea abies* Karst L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. *BMC Genomics* 13: 589.
- Delseny, M., B. Han and Y. Hsing (2010) High throughput DNA sequencing: The new sequencing revolution. *Plant Sci.* 179: 407–422.
- DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna *et*

- al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43: 491–498.
- Doyle, J.J. and J.L. Doyle (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19: 11–15.
- Fekih, R., H. Takagi, M. Tamiru, A. Abe, S. Natsume, H. Yaegashi, S. Sharma, H. Kanzaki, H. Matsumura, H. Saitoh *et al.* (2013) MutMap+: genetic mapping and mutant identification without crossing in rice. *PLoS ONE* 8: e68529.
- Futamura, N., Y. Totoki, A. Toyoda, T. Igasaki, T. Nanjo, M. Seki, Y. Sakaki, A. Mari, K. Shinozaki and K. Shinohara (2008) Characterization of expressed sequence tags from a full-length enriched cDNA library of *Cryptomeria japonica* male strobili. *BMC Genomics* 9: 383.
- Futamura, N., M. Saito, Y. Higuchi, K. Watanabe and K. Shinohara (2011) Screening of genes involved in pollen development of *Cryptomeria japonica* using microarray analyses for comparison of sterile and fertile male strobili derived from pedigrees of four male-sterile mutants. *Jpn. For. Soc. Congr.* 123: Pb198–Pb198.
- Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng *et al.* (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* 29: 644–652.
- Hill, J.T., B.L. Demarest, B.W. Bisgrove, B. Gorski, Y.C. Su and H.J. Yost (2013) MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. *Genome Res.* 23: 687–697.
- Hizume, M., T. Kondo, F. Shibata and R. Ishizuka (2001) Flow cytometric determination of genome size in the Taxodiaceae, Cupressaceae *sensu stricto* and Sciadopityaceae. *Cytologia* 66: 307–311.
- Koboldt, D.C., Q. Zhang, D.E. Larson, D. Shen, M.D. McLellan, L. Lin, C.A. Miller, E.R. Mardis, L. Ding and R.K. Wilson (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22: 568–576.
- le Provost, G., R. Herrera, J.A. Paiva, P. Chaumeil, F. Salin and C. Plomion (2007) A micromethod for high throughput RNA extraction in forest trees. *Biol. Res.* 40: 291–297.
- Li, H. and R. Durbin (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.
- Mishima, K., T. Hirao, M. Tsubomura, M. Tamura, M. Kurita, M. Nose, S. Hanaoka, M. Takahashi and A. Watanabe (2018) Identification of novel putative causative genes and genetic marker for male sterility in Japanese cedar (*Cryptomeria japonica* D. Don). *BMC Genomics* 19: 277.
- Miyajima, D., E. Yoshii, Y. Hosoo and H. Taira (2010) Cytological and genetic studies on male sterility in *Cryptomeria japonica* D. Don (Shindai 8). *J. Jpn. For. Soc.* 92: 106–109.
- Moriguchi, Y., T. Ujino-Ihara, K. Uchiyama, N. Futamura, M. Saito, S. Ueno, A. Matsumoto, N. Tani, H. Taira, K. Shinohara *et al.* (2012) The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in *Cryptomeria japonica* D. Don. *BMC Genomics* 13: 95.
- Moriguchi, Y., K. Uchiyama, S. Ueno, T. Ujino-Ihara, A. Matsumoto, J. Iwai, D. Miyajima, M. Saito, M. Sato and Y. Tsumura (2016) A high-density linkage map with 2560 markers and its application for the localization of the male-sterile genes *ms3* and *ms4* in *Cryptomeria japonica* D. Don. *Tree Genet. Genomes* 12: 57.
- Moriguchi, Y., S. Totsuka, J. Iwai, A. Matsumoto, S. Ueno and Y. Tsumura (2017) Pyramiding of male-sterile genes in *Cryptomeria japonica* D. Don with the aid of closely linked markers. *Tree Genet. Genomes* 13: 61.
- Newton, C.R., A. Graham, L.E. Heptinstall, S.J. Powell, C. Summers, N. Kalsheker, J.C. Smith and A.F. Markham (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Res.* 17: 2503–2516.
- Peterson, B.K., J.N. Weber, E.H. Kay, H.S. Fisher and H.E. Hoekstra (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* 7: e37135.
- Piskol, R., G. Ramaswami and J.B. Li (2013) Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* 93: 641–651.
- Plomion, C., J. Bartholome, I. Lesur, C. Boury, I. Rodriguez-Quilon, H. Lagravelle, F. Ehrenmann, L. Bouffier, J.M. Gion, D. Grivet *et al.* (2016) High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). *Mol. Ecol. Resour.* 16: 574–587.
- Puritz, J.B., C.M. Hollenbeck and J.R. Gold (2014) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2: e431.
- Saito, M. (2010) Breeding strategy for the pollinosis preventive cultivars of *Cryptomeria japonica* D. Don. *J. Jpn. For. Soc.* 92: 316–323.
- Song, K., L. Li and G. Zhang (2016) Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Sci. Rep.* 6: 35736.
- Taira, H., H. Teranishi and Y. Kenda (1993) A case study of male sterility in sugi (*Cryptomeria japonica*). *J. Jpn. For. Soc.* 75: 377–379.
- Taira, H., M. Saito and Y. Furuta (1999) Inheritance of the trait of male sterility in *Cryptomeria japonica*. *J. For. Res.* 4: 271–273.
- Takagi, H., A. Uemura, H. Yaegashi, M. Tamiru, A. Abe, C. Mitsuoaka, H. Utsushi, S. Natsume, H. Kanzaki, H. Matsumura *et al.* (2013) MutMap-Gap: whole-genome resequencing of mutant F2 progeny bulk combined with *de novo* assembly of gap regions identifies the rice blast resistance gene *Pii*. *New Phytol.* 200: 276–283.
- Tsubomura, M., M. Kurita and A. Watanabe (2016) Determination of male strobilus developmental stages by cytological and gene expression analyses in Japanese cedar (*Cryptomeria japonica*). *Tree Physiol.* 36: 653–666.
- Uchiyama, K., T. Ujino-Ihara, S. Ueno, Y. Taguchi, N. Futamura, K. Shinohara and Y. Tsumura (2012) Single nucleotide polymorphisms in *Cryptomeria japonica*: their discovery and validation for genome mapping and diversity studies. *Tree Genet. Genomes* 8: 1213–1222.
- Ueno, S., Y. Moriguchi, K. Uchiyama, T. Ujino-Ihara, N. Futamura, T. Sakurai, K. Shinohara and Y. Tsumura (2012) A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, *Cryptomeria japonica*. *BMC Genomics* 13: 136.
- Ueno, S., Y. Nakamura, M. Kobayashi, S. Terashima, W. Ishizuka, K. Uchiyama, Y. Tsumura, K. Yano and S. Goto (2018) TodoFirGene: Developing transcriptome resources for genetic analysis of *Abies sachalinensis*. *Plant Cell Physiol.* 59: 1276–1284.
- Untergasser, A., H. Nijveen, X. Rao, T. Bisseling, R. Geurts and J.A. Leunissen (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* 35: W71–74.
- Utsunomiya, Y.T., R.V. Alonso, A.S. Carmo, F. Campagnari, J.A. Vinsintin and J.F. Garcia (2013) mendelFix: a Perl script for checking Mendelian errors in high density SNP data of trio designs. *ArXiv*: 1306.2243.
- Velculescu, V.E., L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, Jr., P. Hieter, B. Vogelstein and K.W. Kinzler (1997)

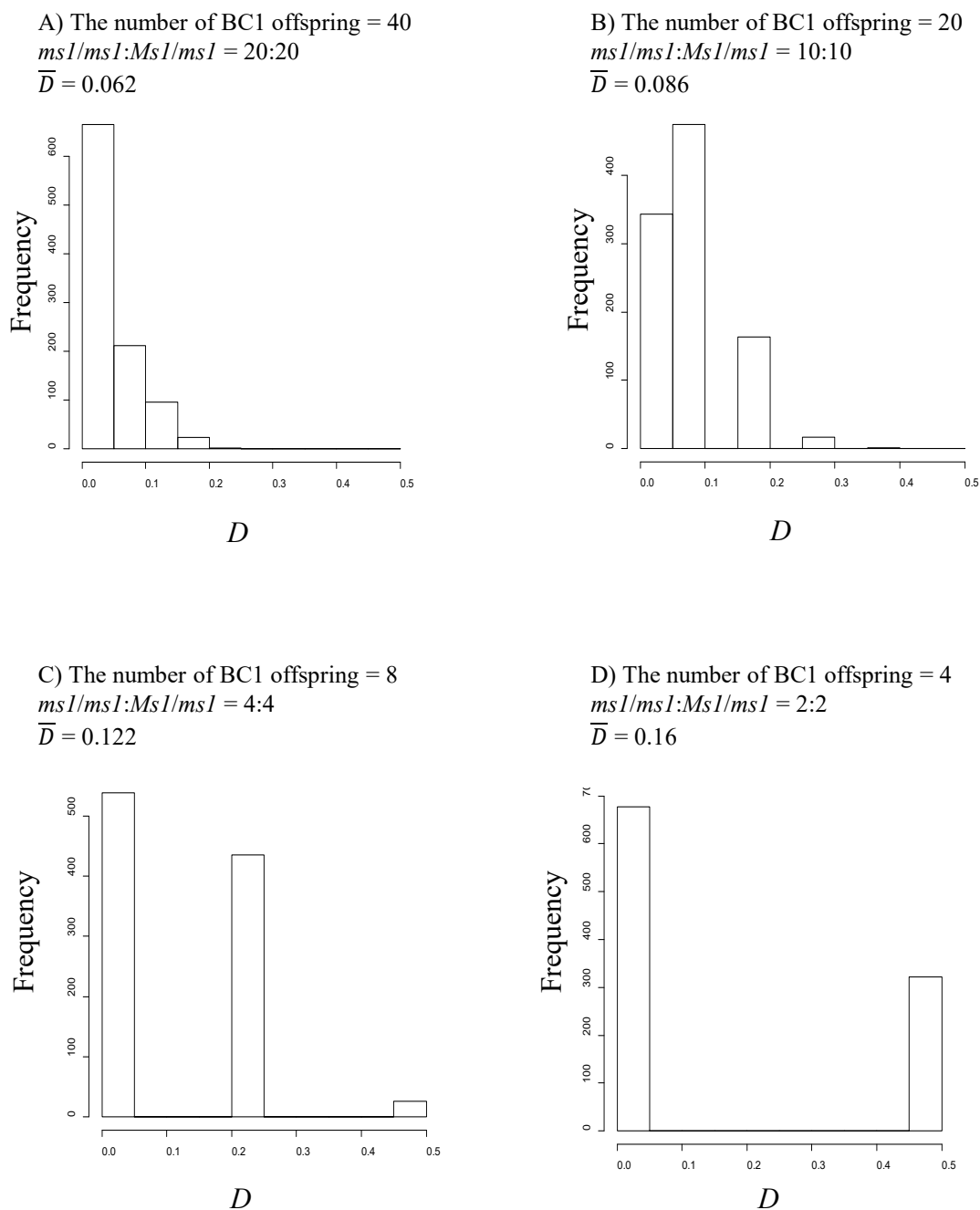
- Characterization of the yeast transcriptome. *Cell* 88: 243–251.
- Voorrips, R.E. (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93: 77–78.
- Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23: 4407–4414.
- Wang, J., M. Lin, A. Crenshaw, A. Hutchinson, B. Hicks, M. Yeager, S. Berndt, W.Y. Huang, R.B. Hayes, S.J. Chanock *et al.* (2009) High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. *BMC Genomics* 10: 561.
- Weber, A.P.M. (2015) Discovering new biology through sequencing of RNA. *Plant Physiol.* 169: 1524–1531.
- Yoshii, E. and H. Taira (2007) Cytological and genetical studies on male sterile Sugi (*Cryptomeria japonica* D. Don), Shindai 1 and Shindai 5. *J. Jpn. For. Soc.* 89: 26–30.



Supplemental Fig. 1.

Schematic representation of selection of candidate reference sequences to calculate *Dav*.

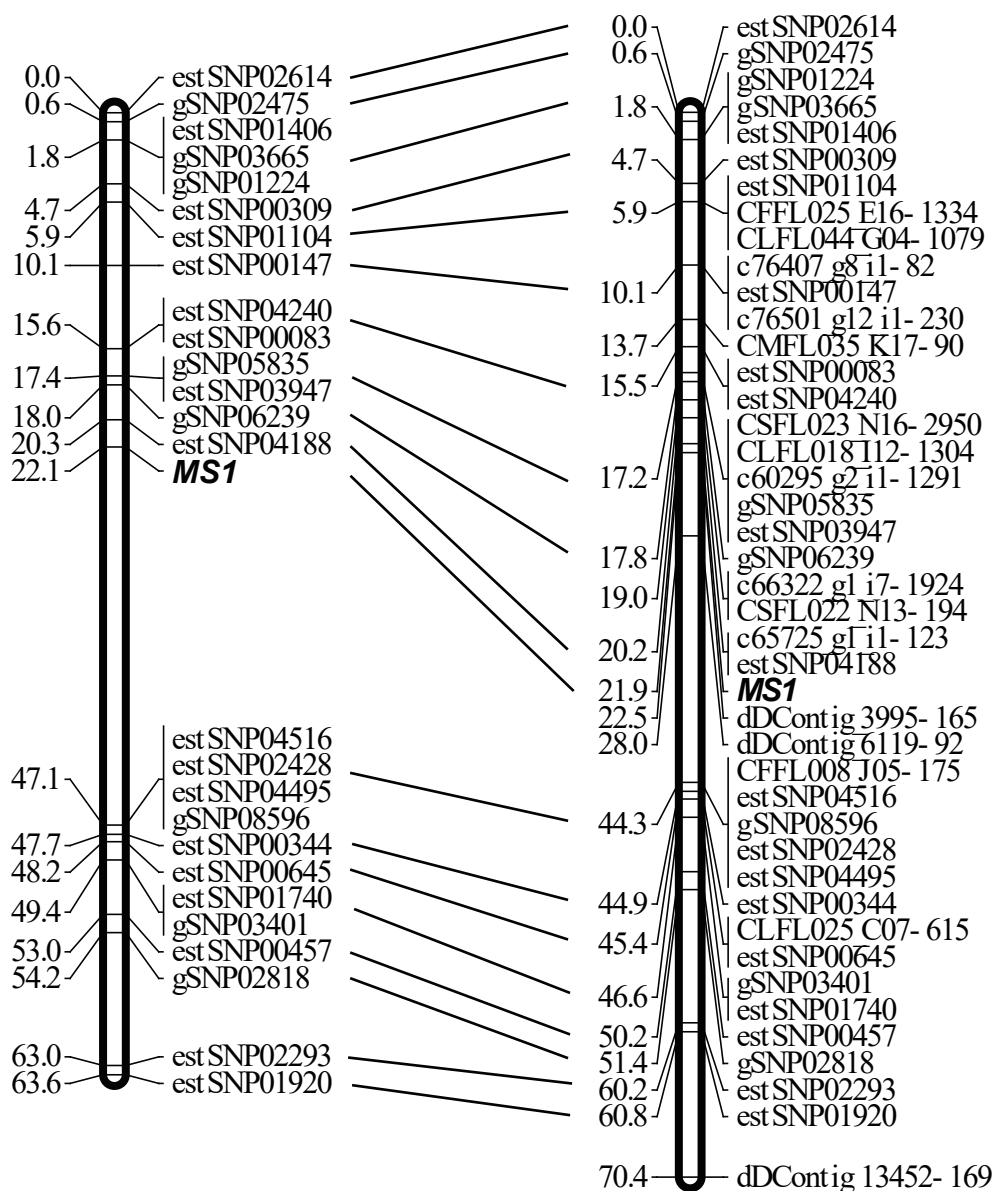
Only the *Dav* was calculated for the reference with at least two SNP sites and with a maximum distance among all SNPs of ≥ 50 bp. This corresponded to about half length of sequencing reads in the present study. Rationale for the condition of maximum distance among SNPs is to avoid multi nucleotide polymorphisms (MNP) which are expected to be under complete linkage and could be considered as a single variant site in the mapping family. If the reference sequences resulted from RAD-Seq, most of them had a stretch of 10 'N's in the middle, which connected forward and reverse reads in paired form. Note that the maximum distance among SNPs are not calculated precisely if the reference had 'N's inside.

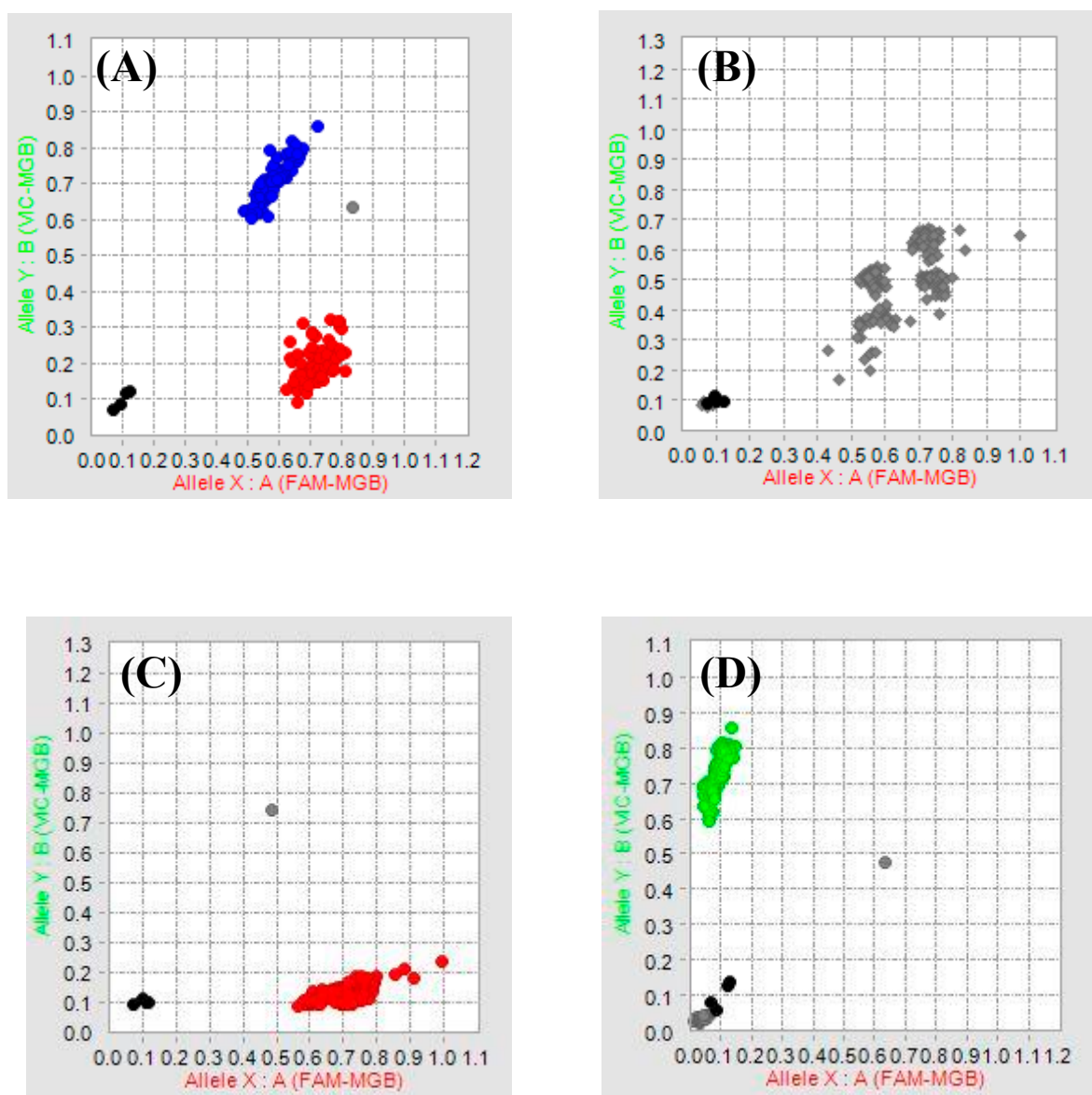


Supplemental Fig. 2.

Simulated distribution of the difference (D)

The simulation was applied in the case of Fig. 1D, where SNP genotypes for parents were XX and XY. According to Mendel's law, XX:XY = 1:1 is expected in the offspring. In the linkage equilibrium, the SNP genotype (XX, XY) is independent of *MSI* genotype (*msl/msl*, *Ms1/ms1*). In this simulation, each BC1 offspring was randomly assigned to male sterile or male fertile phenotype and the D value was calculated. This trial was repeated for 1,000 times, and the histogram of D was drawn and $\bar{D}$ (average of D) was calculated.

**Supplemental Fig. 3.**Improvement of the linkage map (LG9) from Moriguchi *et al.* (2016)

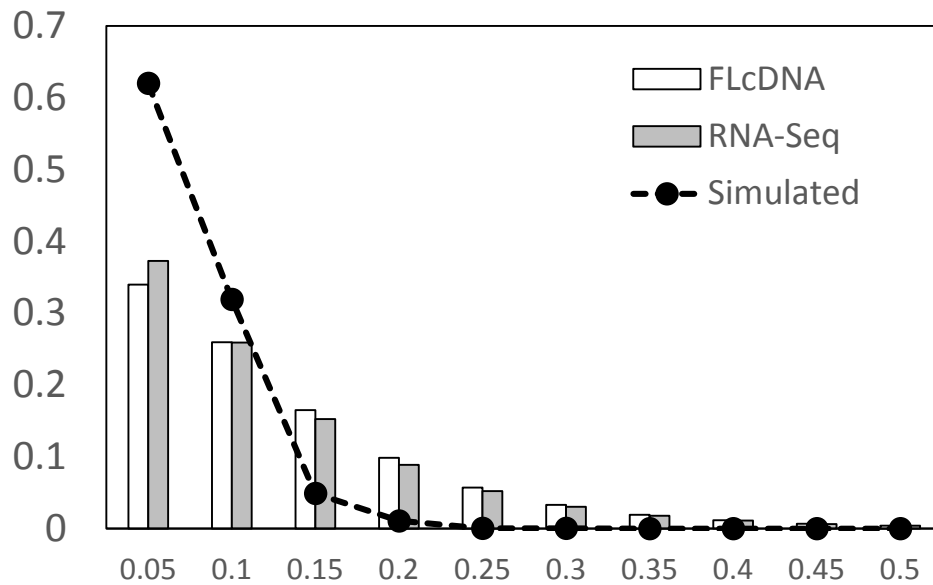


Supplemental Fig. 4.

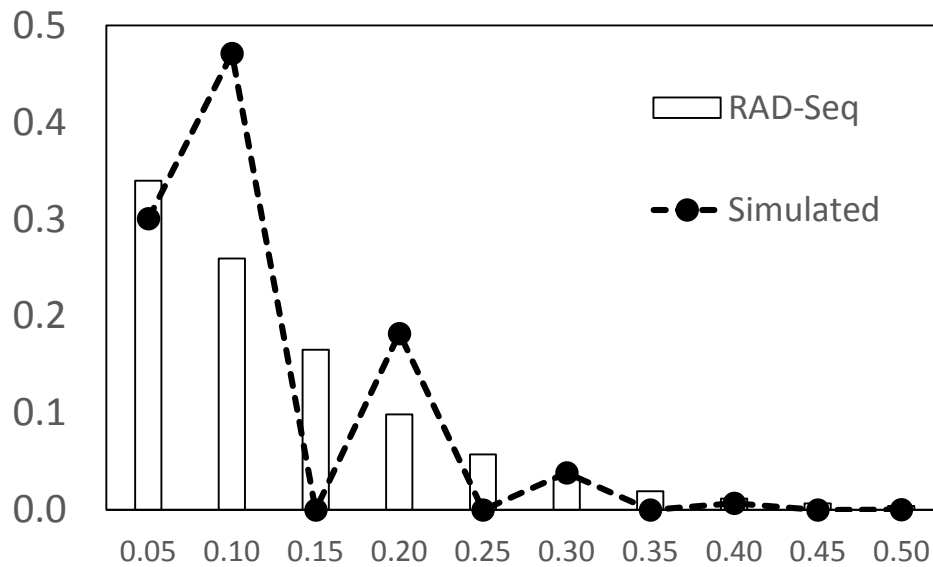
Representative genotyping scatter plots

Displayed are normal (A: dDContig_3995-165) and crushed (B: CFFL034_N06-1783) segregating pattern, and possible monomorphic markers (C: CLFL038_J22-2131, and D: dDContig_15828-101) in the mapping family. Red, blue and green dots indicate XX, XY, and YY genotype, respectively. Black and gray dots indicate negative control and no call samples, respectively.

(A) Histogram for D from FLcDNA and RNA-Seq SNP sites



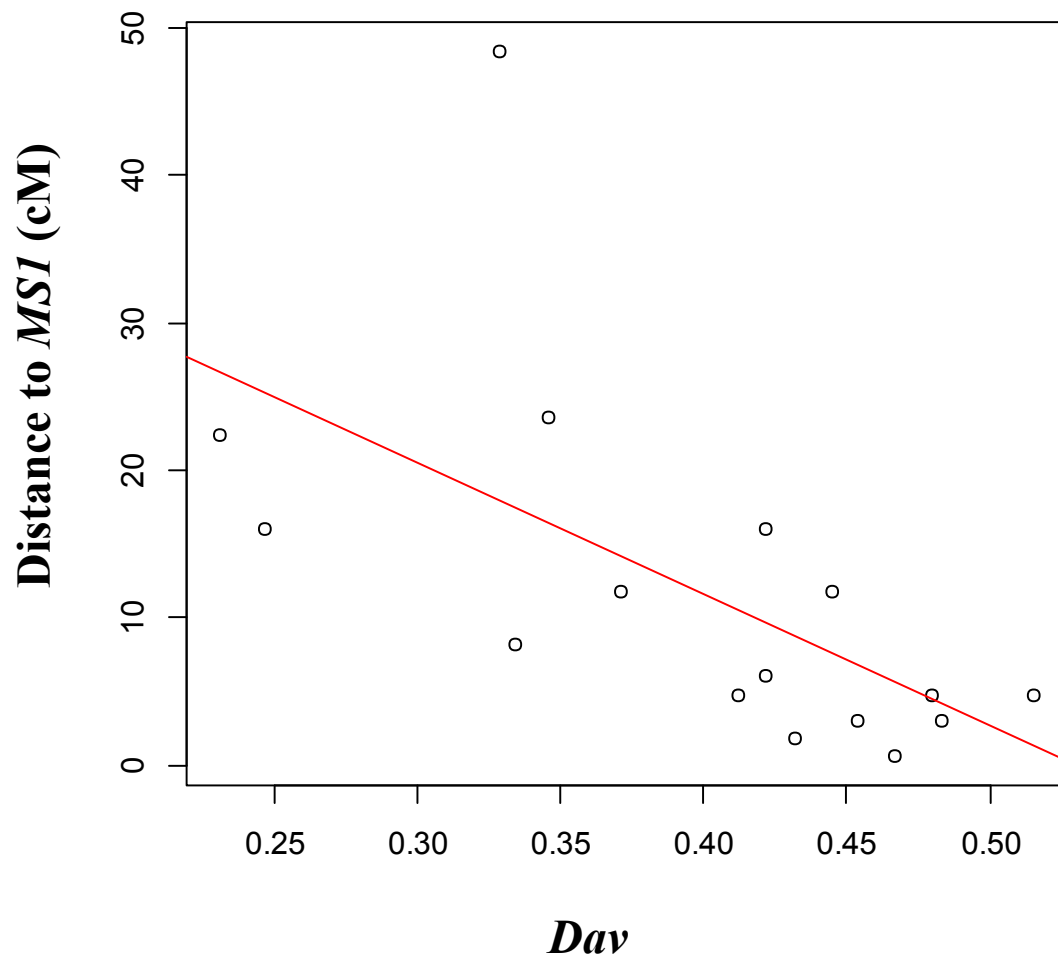
(B) Histogram for D from RAD-Seq SNP sites



Supplemental Fig. 5.

Histogram of the difference (D) for observed/simulated values

The simulation was applied in the case of Fig. 1A, Fig. 1C and Fig. 1D with equal frequency, following the same procedure as in Supplemental Fig. 2. The simulation results were combined and the frequency for each bin was calculated. In this simulation, the number of BC1 offspring (N) was 20 (10 for sterile and 10 for wild type) and 100 (50 for sterile and 50 for wild type), which correspond to cases for FLcDNA and RNA-Seq SNPs and the case for RAD-Seq SNPs, respectively.



Supplemental Fig. 6.

Regression of the distance to *MS1* on *Dav*

The coefficients of the model was estimated with R (R core team 2015), which showed the coefficients of 47.34 and -89.35 for intercept and *Dav*, respectively, and the F-statistic: 8.337 on 1 and 14 degree of freedom with *P* value of 0.01194.

Supplemental Text 1.

DDBJ accession numbers used in the main text.

A: Sequence reads used for mapping to reference: DRR110574-DRR110583, DRR118844-DRR118887

B: Reference sequence by Futamura et al. (2008 and unpublished)

AB359027-AB359028, AB359030-AB359031, AB359033-AB359034, AB359037-AB359038, AB602859, AB728802-AB728803, AB728806, AB728811, AB728814, AB728817, AB728822, AB728824, AB847304, AK406520-AK406552, AK406554-AK406599, AK406601-AK406630, AK406632-AK406691, AK406693-AK406701, AK406703-AK406729, AK406731-AK406734, AK406736-AK406746, AK406748-AK406755, AK406757-AK406778, AK406780-AK406782, AK406785-AK406789, AK406791-AK406794, AK406796-AK406812, AK406814-AK406984, AK406986-AK407015, AK407017-AK407030, AK407032-AK407042, AK407044-AK407049, AK407051-AK407054, AK407056, AK407058, AK407060-AK407062, AK407064-AK407072, AK407074-AK407095, AK407097-AK407098, AK407100, AK407102-AK407155, AK407158-AK407168, AK407170-AK407259, AK407261-AK407266, AK407268-AK407286, AK407288-AK407295, AK407297, AK407299-AK407332, AK407334-AK407339, AK407341-AK407372, AK407374-AK407391, AK407393, AK407395-AK407412, AK407414-AK407424, AK407426-AK407442, AK407444-AK407491, AK407493-AK407497, AK407499-AK407535, AK407537-AK407540, AK407542-AK407568, AK407570-AK407625, AK407627-AK407652, AK407654-AK407658, AK407660-AK407699, AK407701-AK407705, AK407707-AK407710, AK407713-AK407757, AK407759-AK407765, AK407767-AK407772, AK407774-AK407787, AK407789-AK407800, AK407802-AK407804, AK407806-AK407864, AK407866-AK407870, AK407872-AK407893, AK407895-AK407905, AK407907-AK407924, AK407926-AK407927, AK407929-AK407939, AK407941-AK407953, AK407955-AK407961, AK407963-AK407980, AK407982-AK407983, AK407985-AK408054, AK408056-AK408072, AK408074-AK408075, AK408077-AK408079, AK408081-AK408093, AK408095-AK408189, AK408191-AK408199, AK408201-AK408214, AK408216-AK408251, AK408253-AK408256, AK408258-AK408274, AK408276-AK408283, AK408285-AK408299, AK408301-AK408320, AK408322-AK408338, AK408340-AK408350, AK408352-AK408374, AK408376-AK408382, AK408384-AK408390, AK408392-AK408437, AK408439, AK408441-AK408458, AK408460-AK408465, AK408467-AK408468, AK408471-AK408474, AK408476-AK408501, AK408503-AK408529, AK408531-AK408536, AK408538-AK408554, AK408556-AK408569, AK408571-AK408583, AK408585, AK408587-AK408589,

AK408591-AK408592, AK408594-AK408631, AK408633, AK408635-AK408670,
AK408672, AK408675-AK408696, AK408698-AK408709, AK408712-AK408714,
AK408716-AK408723, AK408725-AK408742, AK408744-AK408748, AK408750-
AK408773, AK408775-AK408790, AK408792-AK408804, AK408806-AK408832,
AK408834-AK408843, AK408845-AK408856, AK408858-AK408877, AK408879-
AK408916, AK408918-AK408935, AK408937-AK408948, AK408950-AK408997,
AK408999-AK409042, AK409044-AK409045, AK409047, AK409049-AK409069,
AK409071-AK409117, AK409119-AK409120, AK409122-AK409130, AK409132-
AK409134, AK409136-AK409149, AK409151-AK409167, AK409171-AK409175,
AK409177-AK409178, AK409180-AK409188, AK409190-AK409194, AK409196-
AK409200, AK409202-AK409210, AK409213, AK409215-AK409250, AK409252,
AK409254-AK409272, AK409274-AK409275, AK409277-AK409291, AK409293-
AK409295, AK409297-AK409321, AK409323, AK409325-AK409336, AK409338-
AK409356, AK409358-AK409393, AK409395-AK409440, AK409442, AK409444-
AK409462, AK409464-AK409478, AK409480-AK409495, AK409497-AK409500,
AK409502-AK409526, AK409528-AK409547, AK409549-AK409568, AK409570-
AK409580, AK409582-AK409591, AK409593-AK409638, AK409641-AK409654,
AK409656-AK409662, AK409664-AK409676, AK409678-AK409679, AK409681-
AK409694, AK409696-AK409702, AK409704, AK409706-AK409741, AK409743-
AK409747, AK409749-AK409762, AK409764-AK409782, AK409784-AK409798,
AK409800-AK409839, AK409841, AK409843-AK409876, AK409878-AK409893,
AK409895-AK409896, AK409898-AK409904, AK409906-AK409933, AK409935-
AK409977, AK409980-AK409981, AK409983-AK410085, AK410087-AK410092,
AK410094-AK410103, AK410105-AK410127, AK410129-AK410144, AK410146-
AK410168, AK410172-AK410184, AK410186-AK410193, AK410195-AK410212,
AK410214-AK410243, AK410245-AK410250, AK410252-AK410289, AK410291-
AK410294, AK410296-AK410326, AK410328-AK410333, AK410335-AK410339,
AK410341-AK410346, AK410348-AK410366, AK410368-AK410373, AK410375,
AK410377-AK410379, AK410381-AK410388, AK410390, AK410392-AK410420,
AK410423-AK410446, AK410448-AK410470, AK410472-AK410473, AK410475-
AK410482, AK410484-AK410485, AK410487-AK410519, AK410521-AK410541,
AK410543-AK410553, AK410555-AK410559, AK410561-AK410663, AK410665-
AK410728, AK410730-AK410749, AK410751-AK410764, AK410766-AK410774,
AK410776-AK410809, AK410811-AK410823, AK410825-AK410837, AK410839-
AK410975, AK410977-AK411082, AK411084-AK411101, AK411103-AK411144,
AK411146-AK411156, AK411159-AK411166, AK411168-AK411217, AK411219-
AK411248, AK411250-AK411257, AK411259-AK411293, AK411295-AK411370,
AK411372-AK411382, AK411384-AK411405, AK411407-AK411426, AK411428-
AK411485, AK411489-AK411496, AK411498-AK411543, AK411545-AK411547,
AK411549-AK411557, AK411559-AK411578, AK411580-AK411669, AK411672-
AK411686, AK411688-AK411692, AK411694-AK411719, AK411721-AK411797,
AK411799-AK411803, AK411805-AK411821, AK411823-AK411842, AK411844-
AK411926, AK411928-AK411960, AK411962-AK411979, AK411981-AK411994,
AK411996-AK412020, AK412022-AK412081, AK412083-AK412094, AK412096-

AK412104, AK412106-AK412115, AK412117-AK412134, AK412136-AK412150, AK412152-AK412174, AK412176-AK412213, AK412215-AK412289, AK412291-AK412324, AK412326-AK412346, AK412348-AK412355, AK412357-AK412418, AK412420-AK412468, AK412470-AK412474, AK412476-AK412481, AK412483-AK412493, AK412495-AK412508, AK412510-AK412538, AK412540-AK412561, AK412563-AK412568, AK412570-AK412584, AK412586-AK412611, AK412613-AK412665, AK412667-AK412722, AK412724-AK412748, AK412750-AK412762, AK412764-AK412874, AK412876-AK412908, AK412910-AK412928, AK412930-AK412941, AK412943-AK412992, AK412994-AK413008, AK413010-AK413020, AK413022-AK413026, AK413028-AK413037, AK413039-AK413060, AK413062-AK413125, AK413127-AK413156, AK413158, AK413160-AK413226, AK413228-AK413264, AK413266-AK413351, AK413353-AK413389, AK413391-AK413421, AK413423-AK413432, AK413434-AK413484, AK413486-AK413520, AK413522-AK413537, AK413539-AK413559, AK413561-AK413572, AK413574-AK413615, AK413617-AK413638, AK413640-AK413659, AK413661-AK413668, AK413670-AK413725, AK413727-AK413741, AK413743-AK413764, AK413766-AK413774, AK413776-AK413798, AK413800-AK413809, AK413811-AK413818, AK413820-AK413858, AK413860-AK413892, AK413894-AK413926, AK413928-AK413984, AK413986-AK413999, AK414001-AK414010, AK414012-AK414030, AK414032-AK414046, AK414048-AK414070, AK414072-AK414095, AK414097-AK414103, AK414105-AK414114, AK414116-AK414131, AK414133-AK414135, AK414137-AK414151, AK414156-AK414214, AK414216-AK414227, AK414229-AK414240, AK414242-AK414269, AK414271-AK414305, AK414307-AK414313, AK414315-AK414392, AK414394-AK414403, AK414405-AK414439, AK414441-AK414473, AK414475-AK414503, AK414505-AK414513, AK414515-AK414516, AK414518, AK414521-AK414547, AK414549-AK414558, AK414560-AK414591, AK414593-AK414630, AK414632-AK414639, AK414641-AK414691, AK414693, AK414695-AK414726, AK414728-AK414761, AK414763-AK414774, AK414776-AK414803, AK414805-AK414815, AK414817-AK414963, AK414965-AK414966, AK414968-AK414978, AK414980, AK414982-AK414999, AK415001-AK415002, AK415004-AK415012, AK415014-AK415069, AK415071-AK415104, AK415106, AK415108-AK415200, AK415202-AK415232, AK415234-AK415247, AK415249-AK415286, AK415288-AK415336, AK415338-AK415374, AK415376-AK415407, AK415409-AK415427, AK415429-AK415460, AK415462-AK415479, AK415481-AK415488, AK415490-AK415531, AK415534-AK415604, AK415606-AK415645, AK415647-AK415676, AK415678-AK415697, AK415699-AK415807, AK415809-AK415840, AK415842-AK415903, AK415905-AK415913, AK415915-AK415929, AK415931-AK415940, AK415942-AK415946, AK415948-AK415997, AK415999-AK416006, AK416008-AK416027, AK416029-AK416034, AK416036-AK416060, AK416062-AK416068, AK416070-AK416094, AK416096-AK416108, AK416110-AK416203, AK416205-AK416208, AK416210-AK416241, AK416243-AK416256, AK416258-AK416299, AK416301-AK416305, AK416307-AK416325, AK416327-AK416337, AK416339-AK416360, AK416362-AK416365, AK416367-AK416368, AK416370-AK416443, AK416446-AK416632, AK416634-AK416687, AK416689, AK416691-

AK416696, AK416698-AK416713, AK416715-AK416741, AK416743-AK416744, AK416747-AK416748, AK416865-AK416866, FX334350-FX334830, FX334832-FX334845, FX334847, FX334849-FX335076, FX335078-FX335089, FX335091-FX335145, FX335147-FX335289, FX335291-FX335570, FX335572-FX335642, FX335644-FX335711, FX335713-FX335721, FX335723-FX335724, FX335726-FX335743, FX335745-FX335902, FX335905-FX336212, FX336214-FX336293, FX336295-FX336367, FX336369-FX336428, FX336430-FX336466, FX336468-FX336622, FX336624-FX336721, FX336723-FX336787, FX336789-FX336869, FX336871-FX336934, FX336936-FX337002, FX337004-FX337190, FX337192-FX337249, FX337251-FX337442, FX337444-FX337482, FX337484-FX337824, FX337826-FX337842, FX337844-FX337877, FX337879-FX337945, FX337947-FX338111, FX338113-FX338124, FX338126-FX338264, FX338266-FX338273, FX338275-FX338606, FX338608-FX338684, FX338686-FX338698, FX338700-FX338934, FX338936-FX338960, FX338962-FX339230, FX339232-FX339306, FX339308-FX339309, FX339311-FX339324, FX339326-FX340061, FX340063-FX340076, FX340078-FX340233, FX340235-FX340276, FX340278-FX340325, FX340327-FX340576, FX340578-FX340603, FX340605-FX340609, FX340611-FX340884, FX340886-FX341214, FX341216-FX341246, FX341248-FX341269, FX341271-FX341355, FX341357-FX341419, FX341421-FX341520, FX341522-FX341540, FX341542-FX341617, FX341619-FX341684, FX341686-FX341796, FX341798-FX341895, FX341897-FX341933, FX341935-FX341953, FX341955-FX342052, FX342054-FX342087, FX342089-FX342107, FX342109-FX342127, FX342129-FX342134, FX342136-FX342253, FX342255-FX342293, FX342295-FX342658, FX342660-FX342683, FX342685-FX342731, FX342733-FX342749, FX342751-FX343193, FX343195-FX343622, FX343624-FX343888, FX343890-FX343904, FX343906-FX343938, FX343940-FX344063, FX344065-FX344084, FX344086-FX344115, FX344117-FX344385, FX344387-FX344475, FX344477-FX344784, FX344786-FX344829, FX344831-FX344860, FX344862-FX344878, FX344880-FX344919, FX344921-FX344954, FX344956-FX344971, FX344973-FX345010, FX345012-FX345074, FX345076-FX345120, FX345122-FX345476, FX345478-FX345633, FX345635-FX346285, FX346287-FX346621, FX346623-FX346632, FX346634-FX346748, FX346750-FX347193

Supplemental Text 2.

Simple R simulation code for Supplemental Fig. 2 in the article entitled ‘Scanning RNA-Seq and RAD-Seq approach to develop SNP markers closely linked to *MALE STERILITY 1* (*MS1*) in *Cryptomeria japonica* D. Don’ by Ueno et al.

```
Nsim<-1000          # set the number of simulation
Nkanen<-20          # set the number of fertile trees
Nfunen<-Nkanen      # the number of fertile and sterile trees is the same

XX<-rep(2, length=Nfunen)  # XX vector holds the number of X allele in sterile trees.
XY<-rep(1, length=Nkanen)  # XY vector holds the number of X allele in fertile trees
caseD<-c(XX, XY)
Darray<-numeric()

for(i in 1:Nsim){
  caseDnew<-sample(caseD)
  freq1<-sum(caseDnew[1:Nfunen])/2/Nfunen
  freq2<-sum(caseDnew[(Nfunen+1):(Nfunen+Nkanen)])/2/Nkanen
  Darray[i]<-abs(freq1-freq2)
}

hist(Darray, main=paste("Number of BC1 offspring = ", Nkanen+Nfunen,
"\n", "ms1/ms1:Ms1/ms1 = ", Nfunen, ":", Nkanen, "\n", "mean(D)=", mean(Darray)), xlab="D",
ylab="Frequency", breaks=seq(0,0.5,0.05))
```

Supplemental Text 3 in the article entitled 'Scanning RNA-Seq and RAD-Seq approach to develop SNP markers closely linked to *MALE STERILITY 1 (MS1)* in *Cryptomeria japonica* D. Don' by Ueno *et al.*

Scanning RNA-Seq with marker sequences of known linkage in *C. japonica*

Methods

To empirically survey the distribution of the *Dav* values (see main text for the definition), we applied a scanning RNA-Seq approach to the marker sequences, the map positions of which have already been determined on the YI mapping family (Moriguchi *et al.* 2012, Moriguchi *et al.* 2016). We targeted genic markers and excluded genomic microsatellite markers. We refer to these targeted marker sequences as YI_ref12 and YI_ref16 from the linkage map for Moriguchi *et al.* (2012) and Moriguchi *et al.* (2016), respectively, and collectively called YI_ref hereafter. There were 2,541 and 3,560 sequences in YI_ref12 and YI_ref16, respectively. In total, 383 sequences were included in common in both YI_refs. Because the number of markers in common was small (15 % and 10.8 % for YI_ref12 and YI_ref16, respectively), these two linkage maps may be regarded as mostly independent. Cleaned reads from sterile and fertile samples (22.1 and 22.5 Gb, respectively) were separately mapped onto one of the YI_refs, and the *Dav* values were calculated, including references with a single SNP site. Significant *Dav* values (*Dav_sig*) were defined here as either 5%, 1% or 0.1% of the upper tail of the *Dav* distribution. Because *MS1* was mapped on LG9, the significant *Dav* values from linkage groups other than LG9 were considered false positives. Although false positives were obtained from markers on LG9, map positions of which are far from *MS1*, these were not considered here.

Results

The *Dav* values were obtained for 771 (30.3 %) and 2,160 (85 %) marker sequences from YI_ref12 and YI_ref16, respectively. The *Dav* distribution and that for each linkage group are shown in Figures ST1 and ST2, respectively. The rate of false positives is summarized in Table ST1 according to three significance values.

References

Moriguchi, Y., T. Ujino-Ihara, K. Uchiyama, N. Futamura, M. Saito, S. Ueno, A. Matsumoto, N. Tani, H. Taira, K. Shinohara *et al.* (2012) The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in *Cryptomeria japonica* D. Don. BMC Genomics 13: 95.

Moriguchi, Y., K. Uchiyama, S. Ueno, T. Ujino-Ihara, A. Matsumoto, J. Iwai, D. Miyajima, M. Saito, M. Sato and Y. Tsumura (2016) A high-density linkage map with 2560 markers and its application for the localization of the male-sterile genes ms3 and ms4 in *Cryptomeria japonica* D. Don. Tree Genet. Genomes 12: 57.

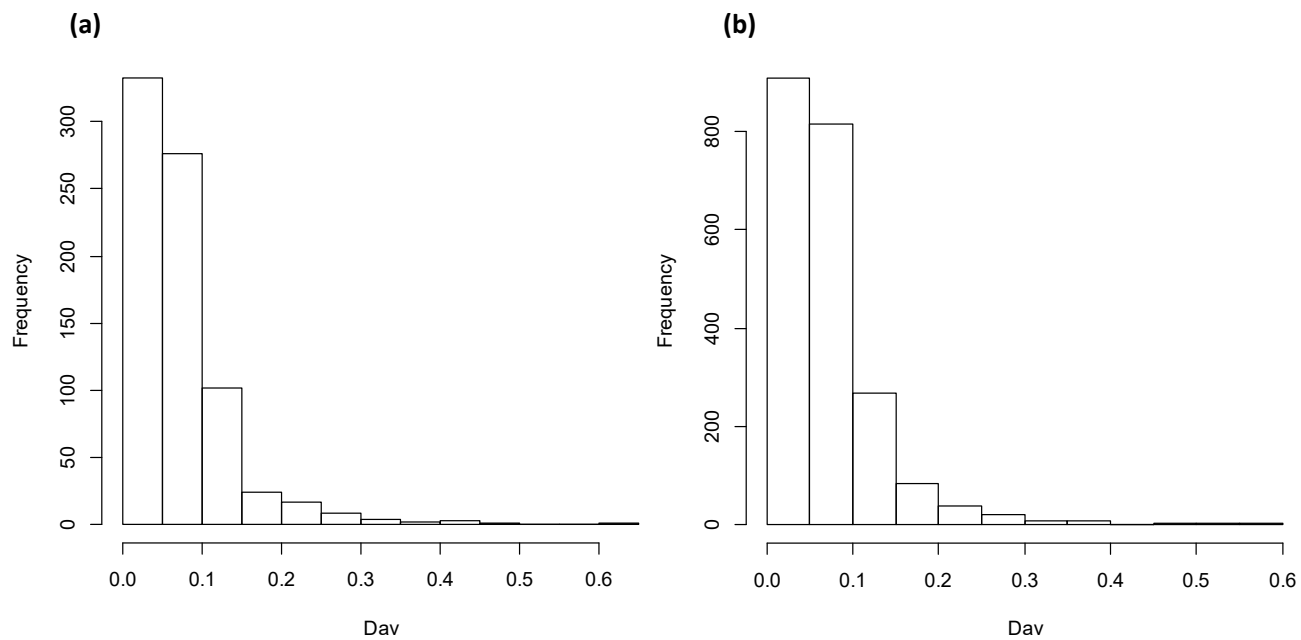


Figure ST1

Dav distribution for the marker sequences from YI_ref12 (a) and YI_ref16 (b)

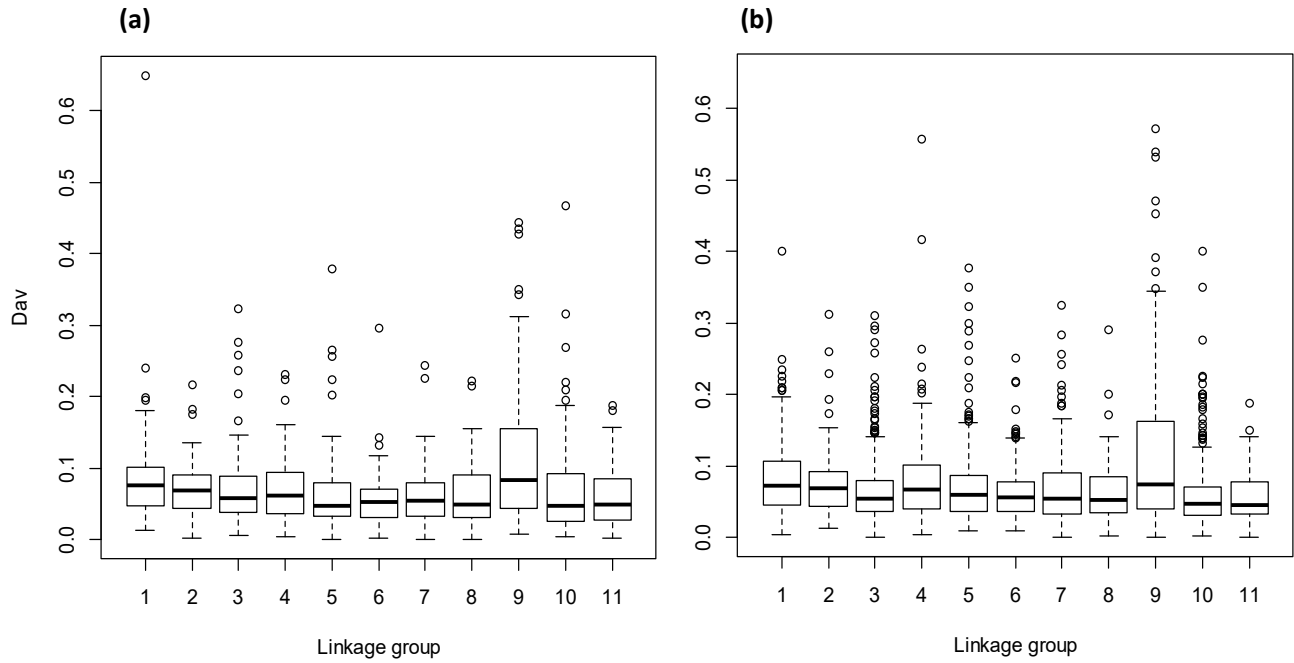


Figure ST2

Dav distribution for YI_ref12 (a) and YI_ref16 (b) on each linkage group for *C. japonica* (2n=22).

MS1 has been mapped on LG9 (Moriguchi *et al.* 2012, Moriguchi *et al.* 2016), where higher *Dav* values are expected.

Table ST1

Summary of significance, number of *Dav* values with significance, *Dav_sig*, number of markers on LG9, and false positives (%) for YI_ref12 (a) and YI_ref16 (b). In total, 771 and 2,160 *Dav* values were considered for marker sequences from (a) and (b), respectively. *MS1* was mapped on LG9, and false positives were defined as markers on linkage groups other than LG9

(a) YI_ref12 (N=771)

Significance (%)	No. <i>Dav</i> with significance	<i>Dav_sig</i>	No. markers on LG9	False positive (%)
0.5	4	0.4354	2	50.00
1	8	0.3432	5	37.50
5	39	0.1958	11	71.79

(b) YI_ref16 (N=2,160)

Significance (%)	No. <i>Dav</i> with significance	<i>Dav_sig</i>	No. markers on LG9	False positive (%)
0.5	11	0.3774	6	45.45
1	22	0.3116	12	45.45
5	108	0.1780	32	70.37