Ann. For. Sci. 66 (2009) 509
© INRA, EDP Sciences, 2009
DOI: 10.1051/forest/2009037

Available online at:
www.afs-journal.org

Original article

Generation of Expressed Sequence Tags and development of microsatellite markers for *Castanopsis sieboldii* var. *sieboldii* (Fagaceae)

Saneyoshi UENO[1]*, Kyoko AOKI[2], Yoshihiko TSUMURA[1]

[1] Tree Genetics Laboratory, Department of Forest Genetics, Forestry and Forest Products Research Institute, 1 Matsunosato, Tsukuba, Ibaraki 305-8687, Japan
[2] Graduate School of Human and Environmental Studies, Kyoto University, Yoshida-nihonmatsu-cho, Sakyo, Kyoto 606-8501, Japan

(Received 22 July 2008; accepted 6 January 2009)

Keywords:
broadleaved tree /
gene ontology /
***Castanopsis cuspidata* var. *cuspidata* /**
***Castanopsis sieboldii* var. *lutchuensis* /**
Castanopsis sieboldii* var. *carlesii

Abstract

• *Castanopsis sieboldii* var. *sieboldii* is an evergreen broadleaved canopy tree that grows on Honshu, Shikoku and Kyushu Islands in Japan. A closely-related, hybridizing, congener species, *C. cuspidata* var. *cuspidata*, has different leaf epidermis structure and different seed nuts morphology compared to *C. sieboldii* var. *sieboldii*. Furthermore, the habitats in which the two species grow often indicate different water requirements.

• Analysis of genetic diversity, using transcribed sequences, will provide a sound basis for the management of this species, allowing successful reforestation, incorporating phylogeographic conservation. In order to obtain transcripts sequences for the species and to develop transcript-based markers, we constructed and analyzed a cDNA library, surveyed microsatellite sequences within it and designed PCR primers for the microsatellites.

• The cDNA library was constructed using tissue from the inner bark of *C. sieboldii* var. *sieboldii*; 3 354 Expressed Sequence Tags (ESTs) were identified. We constructed 2 417 putative unigenes and assigned putative functions to 1 856 of them. Three hundred and fourteen microsatellites were found within the putative unigenes and 16 EST-SSR (Simple Sequence Repeat) markers were developed. The EST-SSR markers developpe in this study should facilitate future analysis of the genetic diversity of this and related species.

Mots-clés :
arbre feuillu sempervirent /
***Castanopsis cuspidata* var. *cuspidata* /**
***Castanopsis sieboldii* var. *lutchuensis* /**
Castanopsis sieboldii* var. *carlesii

Résumé – Generation de marqueurs de séquences exprimées et développement de marqueurs microsatellites pour *Castanopsis sieboldii* var. *sieboldii* (Fagaceae).

• *Castanopsis sieboldii* var. *sieboldii* est un feuillu sempervirent qui pousse sur les îles japonaises de Honshu, Shikoku et Kyushu. *C. cuspidata* var. *cuspidata* une espèce congénère, étroitement liée, hybridée, a une structure de l'épiderme des feuilles différente et une morphologie différente des noix semences par rapport à *C. sieboldii* var. *sieboldii*. En outre, les habitats dans lesquels les deux espèces poussent indiquent souvent des besoins en eau différents.

• L'analyse de la diversité génétique, en utilisant des séquences transcrites, fournira une base solide pour la gestion de cette espèce, ce qui permettra le succès des reboisements, en intégrant une conservation phylogéographique. Afin d'obtenir les transcriptions des séquences de l'espèce et de mettre au point des marqueurs à base de transcription, nous avons construit et analysé une banque d'ADNc, étudié des séquences microsatellites et conçu des amorces PCR pour les microsatellites.

• La banque d'ADNc a été construite en utilisant les tissus de l'écorce interne de *C. sieboldii* var. *sieboldii*, 3 354 marqueurs de séquences exprimées (EST) ont été identifiés. Nous avons construit 2 417 unigènes putatifs et assigné des fonctions putatives à 1 856 d'entre eux. Trois cents quatorze microsatellites ont été trouvés à l'intérieur des unigènes putatifs et 16 EST-SSR (Simple Sequence Repeat) marqueurs ont été développés. Les marqueurs EST-SSR développés dans cette étude devraient faciliter l'analyse future de la diversité génétique de cette espèce et des espèces apparentées.

* Corresponding author: saueno@ffpri.affrc.go.jp

1. INTRODUCTION

Castanopsis sieboldii (Makino) Hatusima var. *sieboldii* is an evergreen broadleaved canopy tree that is found on Honshu, Shikoku and Kyushu Islands in Japan (Yamazaki and Mashiba, 1987b). It is considered to be a climax species in evergreen forests and, in some places, it coexists with its congeneric species *C. cuspidata* (Thunb.) Schottky var. *cuspidata*. However, *Castanopsis sieboldii* var. *sieboldii* is mainly found growing in costal regions, while *C. cuspidata* var. *cuspidata* predominates in drier hilly regions of the interior (Yamanaka, 1966). The two species can be morphologically distinguished by their nut size and shape and the structure of the leaf epidermis (Kobayashi and Sugawa, 1959; Yamazaki and Mashiba, 1987a). *C. cuspidata* var. *cuspidata* has small, globular nuts and a single layer of epidermis cells, while *C. sieboldii* var. *sieboldii* has large, oblong nuts and two layers of epidermis cells. Morphologically intermediate types have been frequently reported, especially at sites where the two species coexist (Kobayashi and Hiroki, 2003; Yamada and Miyaura, 2003). The differences in the distribution and morphology of these two species may reflect natural selection pressures.

The analysis of a large number of genetic markers within gene regions is an effective strategy for investigating the genetic basis of adaptation (Eveno et al., 2008; Kane and Rieseberg, 2007; Tsumura et al., 2007). The genetic information thus acquired can provide valuable insights into various aspects of phylogeography and forest ecology through the study of gene flow, and is also useful in practical applications such as the delineation of seed zones for reforestation programs.

For non-model organisms, one of the most cost-effective methods for obtaining information on genic sequences is to collect and analyze Expressed Sequence Tags (ESTs); these can be obtained by randomly sequencing cloned cDNA derived from mRNA. They represent expressed regions of the genome and can be easily analyzed in single-pass reads using high-throughput capillary sequencers. Another important feature of ESTs is that their putative functions can often be identified by similarity searches against publicly available databases; such functional information may help interpret the results of population genetic studies, particularly with respect to adaptive variations. In fact, some DNA sequences originating from ESTs show signs of adaptive evolution (Kado et al., 2003). As well as being important in fundamental genetic studies, the relationships between genetic variation and adaptation are important to consider in conservational phylogeography.

In the present study, we constructed a cDNA library and generated ESTs using tissue from the inner bark of *Castanopsis sieboldii* var. *sieboldii* to obtain transcribed sequences and to develop transcript-based markers. Putative unigenes were constructed from the ESTs we obtained and the function of the putative unigenes was proposed on the basis of similarity searches against public databases. PCR primers were designed for microsatellites, or simple sequence repeats (SSRs), detected within our database. The ease of use and the highly polymorphic nature of microsatellite markers means that the EST-SSR markers developed in the present study should be useful in genetic diversity studies not only of *C. sieboldii* var. *sieboldii*, but also for related species.

2. MATERIALS AND METHODS

2.1. Source of mRNA and cDNA sequencing

RNA was extracted from several twigs (about 3 cm in diameter) cut from a *C. sieboldii* var. *sieboldii* tree (diameter at breast height ≈ 40 cm) growing in the arboretum at Forestry and Forest Product Research Institute, Tsukuba, Japan, in a fine morning of May 2005. There was no direct sunshine in the morning, because it located at northern side of the institutional building. The origin of the tree was unclear. We used a cutter to peel off the outer bark and then strip the inner bark, which was immediately frozen in liquid nitrogen. The inner bark (ca. 50 g), including the cambium and phloem, was then sliced and ground to a fine powder in liquid nitrogen using a mortar and pestle. The CTAB method (Chang et al., 1993) was used to extract total RNA from the powder; this was further purified using an SV total RNA isolation system (Promega, Madison, USA). The mRNA was recovered using an Oligotex-dT30 Super mRNA purification Kit (Takara, Otsu, Japan); a cDNA Library construction Kit (Stratagene, La Jolla, USA) was used to synthesize first-strand cDNA using oligo $(dT)_{18}$ primers. Synthesized cDNAs were size-selected using CHROMA-SPIN+TE-1000 (Clontech, Mountain View, USA), ligated into pBluescript II SK(+) vectors (Stratagene, La Jolla, USA) and transformed into competent DH10B *Escherichia coli* cells by electroporation. The primary library size was 2.2×10^7 recombinants. The lengths of the insert were checked by PCR with RV-M and M13-47 primers (Takara, Otsu, Japan) using 16 clones; it was confirmed that the average length of the inserts was about 1 700 bp long. After overnight incubation on LB media, cloned inserts in randomly selected white colonies were subjected to rolling circle amplification and sequenced from the 5′ end with dye terminator chemistry using a MegaBACE4000 sequencer (GE Healthcare, Little Chalfont, UK) and T3 primers (Takara, Otsu, Japan). The cDNA library construction and sequencing was performed by the Dragon Genomics Center (Yokkaichi, Japan).

2.2. Analysis of EST sequences

To identify putative functions of the ESTs obtained, we applied the same analytical procedures used for the ESTs identified from *Quercus mongolica* var. *crispula* by Ueno et al. (2008). Briefly, trace2dbest, part of the PartiGene computer package (Parkinson et al., 2004), was used to clean up the raw traces, using the Phred error probability (Ewing and Green, 1998; Ewing et al., 1998), which was set at 0.05. We discarded sequences less than 150 bp long and did not subject them to any further analysis. The remaining sequences were grouped, using CLOBB (Parkinson et al., 2002), into clusters based on similarity. Sequences in the same cluster were then assembled into contigs by means of Phrap (Green, 1999) (using the default parameters in PartiGene). We assumed that all singletons (sequences that were not clustered with other sequences by CLOBB) and contigs (sequences produced by PartiGene) were potential unigenes; these were used for primary annotation via similarity searches. The Blastx (Altschul et al., 1990) software was used to interrogate the NCBI nr database with the e-value cutoff set to 1e-5. Blast was also used to interrogate

the uniprot (uniprot_trembl and uniprot_sprot) databases (Apweiler et al., 2004) to determine a functional classification of the potential unigenes; in this case the e-value cutoff was set at 1e-25. Annot8r_blast2GO, a script developed by Schmid and Blaxter (http://www.nematodes.org/PartiGene/index.html), was used for the annotation. In addition, conserved protein domains were searched against pfam profiles (Finn et al., 2008) using the hmmer ver. 2.3.2 software packages (http://hmmer.janelia.org/), which employ the Hidden Markov Model (HMM).

In order to identify transcripts related to cell wall formation, a similarity search was carried out, using the Blast algorithm against the MAIZEWALL database (Guillaumie et al., 2007), the Cell Wall Navigator database (Girke et al., 2004) and sequences listed by Raes et al. (2003); the e-value cutoff was set at 1e-25.

Characterization of the cDNA library was carried out by the method of Ewing et al. (1999) using a *Populus* EST resource (Sterky et al., 2004) as a reference. The reference *Populus* EST sequences were downloaded from the project web site (http://www.populus.db.umu.se/). Eighteen cDNA libraries (excluding partially subtracted library Y) was used. Additional three *Populus* cDNA library sequences (Leple et al., unpublished; Nanjo et al., 2003) were downloaded from dbEST. Two library sequences by Leple et al. were originated from sivere and mild draught stress-treated mature leaves, while one library sequences by Nanjo et al. (2003) were mixtures from various stress-treated leaves. *C. sieboldii* contigs with more than five ESTs were blasted against each *Populus* cDNA library. The number of hits with score > 200 was counted for each *C. sieboldii* contig. The Pearson correlation coefficient between the number of EST sequences and the number of blast hits for each contig was calculated. The Euclidean distance between two sets was computed as in Ewing et al. (1999).

The microsatellite search tool, SSRIT (Temnykh et al., 2001), from the USDA-ARS Center for Bioinformatics, was used to screen for microsatellite sequences, screening sequences with at least nine, six and five repeats for di-, tri- and tetra-SSRs, respectively. Mono-SSRs did exist in our putative unigenes, however they were not screened here. ESTScan software (Iseli et al., 1999) was used to estimate the location of microsatellites within ESTs (either coding or non-coding regions). Randomization procedures were applied to assess whether the numbers of microsatellite-containing potential unigenes in specific GO categories were significantly overrepresented. Samples of 108 (the number of potential unigenes with microsatellites in all GO categories) were randomly selected from the 1 263 GO-annotated potential unigenes 1 000 times, and 95% confidence limits for the frequency of putative unigenes in each GO category, were determined. This was achieved using Perl scripts written in-house.

2.3. Development of EST-SSR markers

PCR primers were designed by Primer3 for di- and tri-SSRs (Rozen and Skaletsky, 2000); read2Marker script (Fukuoka et al., 2005) was used to automate this process using its default settings. The utility of the EST-SSR primers designed in the present study was estimated by analyzing the polymorphisms among: 10 individuals of *C. sieboldii* var. *sieboldii*, one *C. sieboldii* var. *lutchuensis*, four *C. cuspidata* var. *cuspidata*, and one *C. cuspidata* var. *carlesii* individual (supplementary Tab. I[1]). *C. sieboldii* var. *sieboldii* and *C. cuspidata* var. *cuspidata* were discriminated on the basis of the cell layers of their leaf epidermis following the method described by Kobayashi and Sugawa (1959); individuals of both of these species were sampled from different populations across their distribution ranges in Japan. *C. sieboldii* var. *lutchuensis* is found only in southern Japan (from Amami-oshima to Iriomote Island) (Yamazaki and Mashiba, 1987a) and we used a sample from Okinawa Main Island. *C. cuspidata* var. *carlesii* is found in Taiwan and Southern China and is a relative of *C. cuspidata* var. *cuspidata* (Delectis Florae Reipublicae Popularis Sinicae Agendae Academiae Sinicae Edita, 1998; Yamazaki and Mashiba, 1987b). After DNA was extracted from each leaf sample using a modified CTAB method (Murray and Thompson, 1980), PCR was carried out in 10 μL reaction mixtures containing ca. 10 ng genomic DNA, 1 × PCR buffer, 200 μM of each dNTP, 1.5 mM $MgCl_2$, 0.2 μM of each primer designed in the present study and 0.4 U of *Taq* polymerase (Promega, Madison, USA), using the following PCR program: 94 °C for 3 min; then 40 cycles of 94 °C for 45 s, 55 °C for 45 s and 72 °C for 45 s, followed by a final extension at 72 °C for 10 min. PCR products were labeled with ChromaTide Rhodamine Green-5-dUTP (Molecular Probes Eugene, USA) according to the method of Kondo et al. (2000), and analyzed using an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, USA). The number of alleles (*Na*) and observed heterozygosity (H_o) (Nei and Kumar, 2000) were determined for each locus.

In order to examine transferability of EST-SSR markers developed here to related species , we carried out cross-spceies PCR amplification for eight Fagaceae species (*Fagus crenata*, *Lithocarpus glabra*, *Castanea crenata*, *Quercus glauca*, *Q. dentata*, *Q. serrata*, *Q. mongolica* and *Q. variabilis*), which naturally occur in Japan. PCR was performed in the same reaction conditions as above. The PCR products were electrophoretically separated on 2% agarose gels and stained with ethidium bromide. The resulting banding pattern was recorded as a single main band (+), no amplification (–) or multiple-banding pattern (m).

The discriminant power of EST-SSR markers for different species and varieties in *Castanopsis* was tested by using eight individuals of *C. sieboldii* var. *sieboldii*, *C. sieboldii* var. *lutchuensis*, *C. cuspidata* var. *cuspidata* and *C. cuspidata* var. *carlesii*. These individuals were from Ibaraki, Okinawa and Hiroshima (Suppl. Tab. I[1]) and Taiwan (Aoki et al., unpublished) population, respectively. PCR was carried out as described before except that fluoresent primer was used. PCR products were analyzed on ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, USA). The proportion of shared alleles between individuals was computed by the MSA software (Dieringer and Schlötterer 2003) and an NJ dendrogram was constructed with the MEGA ver. 4.0.2 (Tamura et al., 2007).

3. RESULTS

3.1. cDNA sequencing and EST analysis

We obtained a total of 3 638 sequences and identified 3 354 high quality readings (DDBJ accession numbers DC600172 to DC603525) that were more than 150 bp long (average length, 552 bp). The high quality sequences were grouped into 423 clusters, which contained 1 386 EST sequences. The remaining 1 968 sequences were singletons. When sequences in the same cluster were assembled into contigs, 449 contigs were created. We identified a total of 2 417 potential unigenes (contigs and singletons). The ratio of unique sequences in the library (= the number of unigenes/the number of high quality

[1] Supplementary material is available at www.afs-journal.org.

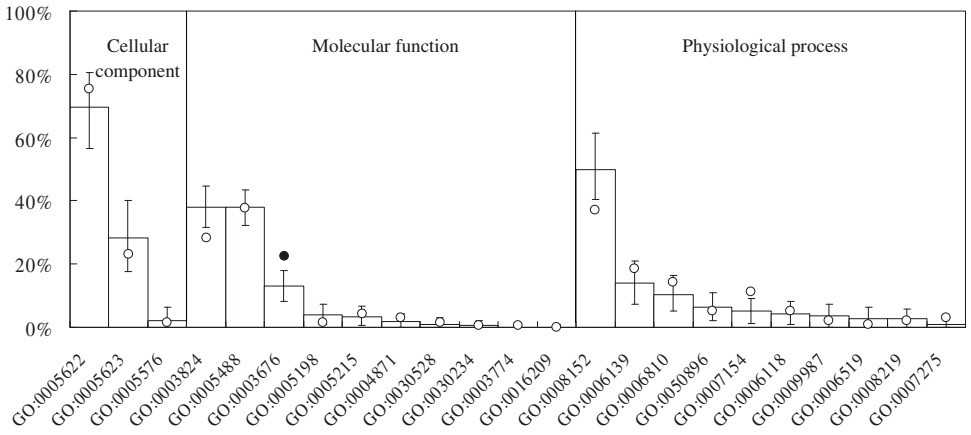

Figure 1. Functional profile of the 1 263 *C. sieboldii* putative unigenes and the 108 microsatellite-containing putative unigenes annotated according to GO slim terms (squares and circles, respectively). The error bars indicate 95% confidence limits for the frequencies of putative unigenes annotated with the specific GO slim term when 108 putative unigenes were randomly sampled 1 000 times. Significantly highly represented GO slim terms amongst putative unigenes with microsatellites are indicated by the closed circle. The GO IDs and corresponding terms are as follows: GO:0005622, intracellular; GO:0005623, cell; GO:0005576, extracellular region; GO:0003824, catalytic activity; GO:0005488, binding; GO:0003676, nucleic acid binding; GO:0005198, structural molecule activity; GO:0005215, transporter activity; GO:0004871, signal transducer activity; GO:0030528, transcription regulator activity; GO:0030234, enzyme regulator activity; GO:0003774, motor activity; GO:0016209, antioxidant activity; GO:0008152, metabolic process; GO:0006139, nucleobase, nucleoside, nucleotide and nucleic acid metabolic process; GO:0006810, transport; GO:0050896, response to stimulus; GO:0007154, cell communication; GO:0006118, electron transport; GO:0009987, cellular process; GO:0006519, amino acid and derivative metabolic process; GO:0008219, cell death; GO:0007275, multicellular organismal development.

sequences) was 0.721. The average number of EST sequences per contig was 3.1, average size of the sequence in the contig was 749 bp. The largest cluster included 39 ESTs, while the other clusters contained less than 30 ESTs (Suppl. Tab. II[1]).

In total, 1 856 potential unigenes had similarities with other proteins in the NCBI nr database at the e-value cutoff level of 1e-5. The remaining 561 potential unigenes, however, did not have any hits with e-value cutoff level of 1e-5. The gene ontology (GO) functional classification assigned 533, 1 100 and 875 of the potential unigenes to the cellular component, molecular function and biological processes categories, respectively. The GO terms represented by the largest numbers of potential unigenes within each of these categories were: intracellular (GO:0005622; 70%), catalytic activity (GO:0003824; 38%), binding (GO:0005488; 38%) and metabolic processes (GO:0008152; 50%) (Fig. 1). At least one GO term was assigned to 1 263 of the potential unigenes. HMM searches against the pfam database recognized 695 putative unigenes at the e-value cutoff level of 1e-10. Four hundred protein families were identified, with the ubiquitin family being the most common (Tab. I). The other common protein families included AP2 domain, RNA recognition motif, DnaJ domain, ubiquitin-conjugating enzyme and core histone H2A/H2B/H3/H4.

The Blastx search against the Cell Wall Navigator database identified 45 putative unigenes with a significant similarity at the e-value cutoff level of 1e-25; of these, 37 were singleton putative unigenes. In addition, the Tblastx search against MAIZEWALL database and the sequences listed by Raes et al. (2003) identified 170 and six putative unigenes, respectively. In total, 193 putative unigenes including 312 transcripts (9.3%) exhibited a similarity to genes for cell wall formation. Examples were callose synthase, cellulose synthase-like (Csl), trans-cinnamate 4-hydroxylase (C4H) and cinnamyl alcohol dehydrogenase 2 (CAD2).

Table I. Abundant protein families (number of putative unigenes > 5).

Pfam protein family	No. of putative unigenes
Ubiquitin family	18
AP2 domain	15
RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	12
DnaJ domain	10
Ubiquitin-conjugating enzyme	10
Core histone H2A/H2B/H3/H4	10
Mitochondrial carrier protein	7
TIR domain	7
Miro-like protein	7
Ras family	7
ADP-ribosylation factor family	7
Myb-like DNA-binding domain	6
Major intrinsic protein	6
Peroxidase	6

Calculation of correlation coeffiient between the number of EST sequences in a contig and the number of blast hit with the contig produced a 61 × 22, contig by library matrix, which were then used to calculated Euclidian distance among cDNA

Table II. Abundance of microsatellites in the *C. sieboldii* putative unigenes.

SSR	Motif	Occurrence	
		Total	Coding
	AG	151	0
	AT	23	0
	AC	9	0
	CG	0	0
Di-SSR	Total	183	0
	AAG	37	19
	AGG	23	21
	AAC	17	14
	ACC	17	15
	AGC	8	8
	AAT	7	0
	ATC	6	5
	CCG	6	6
	ACT	3	2
	ACG	1	1
Tri-SSR	Total	125	91
	GAGC	1	0
	GTCA	1	0
	TTAT	1	0
	TCTT	1	0
	ACCA	1	0
	TCTT	1	0
Tetra-SSR	Total	6	0
Total		314	91

libraries. The cDNA library in the present study was the closest to Leple et al. library which was originated from sivere draught stress-treated mature leaves. Libraries of mild draught stress-treated mature leaves by Leple et al, wood cell death (X) and senescing leaves (I) by Strerky et al. (2004) were also closer than the other libraries (Suppl. Tab. III[1]).

We detected 314 microsatellites in 2 417 potential unigenes (Suppl. Tab. IV[1]), with some potential unigenes containing up to three microsatellites. The numbers and relative proportions of di-, tri- and tetra-SSRs were 183 (58.3%), 125 (39.8%) and 6 (1.9%), respectively. The AG motif was the most common, while there were no CG repeats in the potential unigenes (Tab. II). Among the tri-SSRs (98% of which had fewer than 10 repeats), AAG and AGG repeats were relatively frequent, however, there were fewer tri-SSRs than di-SSRs. All six of the tetra-SSRs had a different motif. The maximum number of repeats was 71 (of the AG motif). Estimation of the locations of the SSRs within the ESTs suggested that 176 (56%) of them reside in non-coding regions, and 91 (29%) of them are likely to be located in coding regions. The locations of the 47 (15%) remaining SSRs could not be determined because of no data returned by ESTScan software, so they are assumed as non-coding in the present study. No di-SSRs were identified as coding, while 91 (72.8%) of the tri-SSRs appeared to be located in coding regions (Tab. II). GO annotations, which were assigned to a total of 108 microsatellite-containing putative unigenes, indicated that GO:0003676 (nucleic acid binding) was strongly overrepresented (22.5%; $P < 0.01$) even after the levels of significance were adjusted by Bonferroni criteria (Fig. 1).

3.2. Analysis of EST-SSR markers

In total, 39 and 24 primer pairs were designed for the di- and tri-SSRs, respectively, although some of the target sequences contained both di- and tri-SSRs. Primers were designed for all sequences with more than or equal to 10 repeat (max) in the read2Marker software output. Twenty-nine sequences were in this category. Additional 34 sequences for primer design were selected arbitrary from repeat (max) more than 6 and less than 9. Thirty-seven of the 63 designed primer pairs amplified *C. sieboldii* var. *sieboldii* genomic DNA within the expected size range. Nine primer pairs produced larger fragments than expected, probably due to the occurrence of introns. These loci are impossible to genotype by capillary sequencers because of the large fragment size; they were excluded from further analysis. The remaining primer pairs had no amplifications or multi-banding pattern. After fluorescent labeling and electrophoresis by the capillary sequencer, 16 primer pairs (Tab. III) showed a clear single locus amplification pattern for 16 of the genomic DNA samples. Some primer pairs produced three peaks and/or peaks that were difficult to genotype; these are not listed in Table III. Four potential unigenes (CcC01513_1, CcC02069_1, CcC02291_1 and CcC02375_1) were predicted, by ESTScan, to have coding SSRs. Most of the markers listed in Tab. III were found within the unigene sequences that had similarities with other proteins in the nr database. The number of alleles per locus (*Na*), and observed heterozygosity (H_o) for *C. sieboldii* var. *sieboldii* ranged from 3 to 9, and 0.2 to 0.9, respectively (Tab. IV). For *C. cuspidata* var. *cuspidata*, *Na* and H_o ranged from 2 to 6, and 0 to 1, respectively (Tab. IV).

When these primers were applied to related Fagaceae species, three (19%) of them amplified all Fagaceae species (Tab. V). When applied to *C. crenata*, 14 (88%) of the primer pairs produced single PCR product. However, for *F. crenata*, the transferability was low, with half of the primre pairs no amplification. Within two *Castanopsis* species and two variaties, all of the markers were well applied. Genotypes were cleary determined for all 32 individuals. The NJ dendrogram using shared allele distance between individuals (Fig. 2) showed clear split between *C. sieboldii* and *C. cuspidata*. However, discrimination of the varieries (*C. sieboldii* var. *lutchuensis* and *C. cuspidata* var. *carlesii*) from either of *C. sieboldii* var. *sieboldii* or *C. cuspidata* var. *cuspidata* was not so clear.

4. DISCUSSION

4.1. Characteristics of putative unigenes

In the present study, a non-normalized cDNA library was constructed from RNA extracted from the inner bark of *C. sieboldii* var. *sieboldii.* The largest cluster obtained contained 39 transcripts (1.2% of the total), indicating the suitability of bark tissue for library construction without normalization. Bhalerao et al. (2003) reported that, with young leaf tissue as the RNA source, 14% of the total number of their

Table III. Characteristics of the *C. sieboldii* EST-SSR markers.

Locus Accession No.	Primer sequence	Repeat region	nr BlastX top hit[#] accession number [species] Description (E-value)
CcC00055	TCCTTCATGTGAGATGTGGGAGT	$(TA)_5AA(TA)_9$	ref|NP_176736.2| [*Arabidopsis thaliana*]
DC600400	GACCTCACCCTCAACCTTGCTCT		Inositol or phosphatidylinositol phosphatase (2E-10)
CcC00087	GGAACAACTTCTACTCCGCCCTC	$(TC)_{13}$	ref|NP_174541.1| [*Arabidopsis thaliana*]
DC600434	CGTTTCCTGGTTGAAGAGAGCAA		ATMYC2; DNA binding / transcription factor (2E-11)
CcC00610	CAGCAAAGAAGGCTGATGTGTCA	$(TA)_{12}TT(TA)_3$	–
DC601033	GCAAGTGTTGAAAGCAAACATGG		
CcC00660	GCCTCAACAGTCATCACAAAACACA	$(TC)_8TT(TC)_4ATAAAC(TC)_3TT(TC)_4$	–
DC601093	GAAACACAAGCACCGATCCAATC		
CcC00972	AATTTCACATCCCAACTCTGCGA	$(AG)_{13}$	gb|AAT08725.1| [*Hyacinthus orientalis*]
DC603076	TGGAGGGAGTAGTGGACGATCAA		Histone H4 (4E-36)
CcC01471	GCACGAGGCTTTGTTCTCCTAGA	$(TC)_3ATTCTT(TC)_{14}$	ref|NP_565678.1| [*Arabidopsis thaliana*]
DC602823	AGACGCAGAGCCAATGAATGAAG		CXIP4 (CAX INTERACTING PROTEIN 4) (8E-44)
CcC01513*	GCAATTCCAGCAACCACAGAATC	$(TCA)_5(GCA)_6ACAGCAACAACAGCAG(CAA)_3CA(GCA)_3AC(AGC)_3$	–
DC602232	AAGGAATTGCTTGCAAGGATGGT		
CcC02014	CACGAGGGAGAAACTTCGCATT	$(TC)_{11}TGTGATCGATCGCCGAGAAA(GAA)_6$	ref|NP_189513.1| [*Arabidopsis thaliana*]
DC602918	TGTAATCAGAGGCGGTGAGAAGC		Hydrogen-transporting ATP synthase (1E-83)
CcC02022	TTCACTTGTTTTTCCCGACCAGA	$(TC)_{12}$	gb|AAS10179.1| [*Antirrhinum majus*]
DC602928	CCGCTAAAATGGTGTTGCAGAAG		YABBY2-like transcription factor YAB2 (3E-26)
CcC02069*	ATCACCGGAGAAAACCCTAACGA	$(CAA)_6$	–
DC602996	AATGTTTCGGACCAATTCGAGGT		
CcC02075	GCACGAGGGTTGTCTTCTAACCT	$(TC)_9$	ref|NP_176576.1| [*Arabidopsis thaliana*]
DC603005	TTATCTGTTGCAGGCATTGCTGA		Unknown protein (5E-45)
CcC02291*	TATCAGCCACAGCTACCGTCACA	$(GCA)_9$	ref|NP_565849.1| [*Arabidopsis thaliana*]
DC603314	AGCGCGGTCTTGTACTGACTCAC		Unknown protein (4E-23)
CcC02375*	TGGGTTAGCCGACAATCATGAAC	$(GAA)_3CA(AGA)_3CAAGAAGATGAAGAGGATTAGAGTC(GAA)_7$	ref|NP_194152.1| [*Arabidopsis thaliana*]
DC603438	CCACCGAGAGCTAAGACCACAGA		SLY1 (SLEEPY1) (1E-41)
CcC02438	GCACGAGGCTTTTTCTCAATCTC	$(AC)_5(TC)_9$	dbj|BAB10064.1| [*Arabidopsis thaliana*]
DC600268	TTGGTTTGTTCACGGAGATCCAC		Unnamed protein product (2E-24)
CcC02464	CACGAGGCTTTTCTTCCTTTCCT	$(TC)_{20}$	ref|NP_849880.1| [*Arabidopsis thaliana*]
DC600310	TTGCTGTTGTTGTTGTTGTTGCC		Acid phosphatase (5E-24)
CcC02535	ACGAGGCAATCCCGTAAAAAGAA	$(TA)_{11}$	ref|NP_197067.2| [*Arabidopsis thaliana*]
DC600239	TGTTCATTGTCACTGCAGCCTGT		COBL4/IRX6 (3E-42)

*: Repeat region predicted, by ESTScan, to be present in the coding region.
[#]–: No hits found in the nr database.

Table IV. Polymorphisms of the EST-SSR markers, based on samples of 10 *C. sieboldii* var. *sieboldii* (CSS), four *C. cuspidata* var. *cuspidata* (CCC), one *C. sieboldii* var. *lutchuensis* (OKN) and one *C. cuspidata* var. *carlesii* (TWR). For OKN and TWR individuals, allele size is shown.

Locus	Expected size	CSS (N = 10)			CCC (N = 4)			OKN	TWR
	(bp)	*Na*	H_o	Size range (bp)	*Na*	H_o	Size range (bp)		
CcC00055	285	5	0.8	285–297	5	0.5	277–293	287/295	281/283
CcC00087	400	5	0.7	391–401	5	0.75	391–401	393/397	399/401
CcC00610	381	7	0.9	370–385	3	1.0	370–376	374/380	374/378
CcC00660	132	3	0.2	133–137	4	0.5	125–141	135/137	133/133
CcC00972	257	8	0.7	241–261	4	0.75	241–253	251/251	243/259
CcC01471	129	8	0.5	110–173	4	0.5	114–125	125/133	121/125
CcC01513	375	3	0.4	368–377	2	0.5	368–374	374/374	368/368
CcC02014	201	7	0.7	201–220	2	0.0	201–203	201/220	201/201
CcC02022	347	9	0.8	352–372	4	0.25	353–365	356/356	355/373
CcC02069	123	6	0.8	117–129	6	0.5	117–136	117/120	114/117
CcC02075	211	5	0.6	214–224	4	0.75	212–224	216/216	206/206
CcC02291	107	3	0.3	94–106	2	0.5	97–100	106/106	100/103
CcC02375	318	6	0.6	309–332	3	0.75	321–327	309/321	324/324
CcC02438	319	7	0.5	271–322	5	1.0	267–283	271/271	275/275
CcC02464	230	8	0.9	206–226	4	0.25	208–218	216/224	220/234
CcC02535	369	7	0.7	365–379	6	0.75	364–375	375/375	373/375

Na: Number of alleles per locus, H_o: Observed heterozygosity.

Table V. Cross-species amplification of 16 EST-SSR primer pairs from *C. sieboldii*. Fc, Lg, Cc, Qg, Qd, Qs, Qm and Qv indicate *Fagus crenata*, *Lithocarpus glabra*, *Castanea crenata*, *Quercus glauca*, *Q. dentata*, *Q. serrata*, *Q. mongolica* and *Q. variabilis*, respectively. The plus (+) and minus (-) signe and 'm' indicate single locus amplification, no products and multiple banding pattern, respectively.

Primers	Fc	Lg	Cc	Qg	Qd	Qs	Qm	Qv
CcC00055	m	–	+	+	+	+	–	m
CcC00087	–	+	+	–	–	m	–	–
CcC00610	–	+	+	–	+	+	+	+
CcC00660	+*	–	+	m	+	+	+	+
CcC00972	–	+	+	+	+	+	+	+
CcC01471	–	+	–	–	–	–	–	–
CcC01513	+	+	+	+	+	+	+	m
CcC02014	+	+	+	+	+	+	+	+
CcC02022	+	+	+	+	+	+	+	+
CcC02069	+	+	+	+	+	+	+	+
CcC02075	–	–	+	m	m	m	–	–
CcC02291	+	+	+	+	+	+	+	m
CcC02375	+	m	+	+	m	+	+	m
CcC02438	–	m	–	m	+	+	+	m
CcC02464	–	+	+	+*	+	+	+	–
CcC02535	–	–	+	–	+	–	+	–

* Products more than 1 kbp.

clones represented the RbcS gene (a small subunit of Rubisco). Although the objectives and method to construct putative unigenes was different among studies, young leaf library by Bhalerao et al. (2003) had 1 943 unique gene families from 4 923 EST sequences, leading to the ratio of unique sequence of 0.395. The ratio shows great contrast to that (0.721) of the present study.

The non-normalized nature of the library was reflected in the sizes of the clusters, which indicated the relative levels of gene expression. The most abundant ESTs (Suppl. Tab. II[1]) included an RD22-like protein (CcC00008_1; 39 ESTs), a metallothionein-like protein (CcC00064_1; 19 ESTs), and a translationally controlled tumor protein (CcC00353_1; 13 ESTs), whose expression is thought to be related to stress (Bommer and Thiele, 2004; Navabpour et al., 2003; Yamaguchi-Shinozaki and Shinozaki, 1993) These transcripts may reflect environmental conditions where the tree grows and/or the active physiological state of the tree. When the library was associated with reference and stress libraries of *Populus*, it was related more to the sivere draught stress-treated library. Other cluster types were related to transcription and translation. The cluster CcC00231 had similarity to translation elongation factor 1α and had hits with ESTs in all the other libraries, probably due to "housekeeping" role of the gene. Translationally controlled tumor proteins (corresponding to the cluster CcC00353) may also be involved in the elongation step of protein synthesis (Cans et al., 2003). The other highly represented clusters (CcC00021 and CcC00068) showed similarities with asparaginase. Rapidly growing tissues (e.g. apical meristems, expanding leaves, inflorescences and seeds) are known to display asparaginase activity (Grant and Bevan, 1994); the cambium (lateral meristem) of the inner bark is one such fast-growing, actively developing tissue. Thus, the detected ESTs that resemble asparaginase are probably related to the lateral growth of the stem, although low temperature stress may also induce asparaginase expression (Cho et al., 2007). In EST analysis of wood-forming tissue of poplar (Sterky et al., 1998), putative genes for cyclophilin, a translationally controlled tumor protein, S-adenosyl-L-methionine synthase, the elongation factor 1-α and a 14-3-3-like protein were identified as the most highly expressed genes. Furthermore, analysis of *Cryptomeria japonica* inner bark EST (Ujino-Ihara et al., 2000) revealed that peptidyl-proryl cis-trans isomerase, a protein translation factor SUI1 homologue,

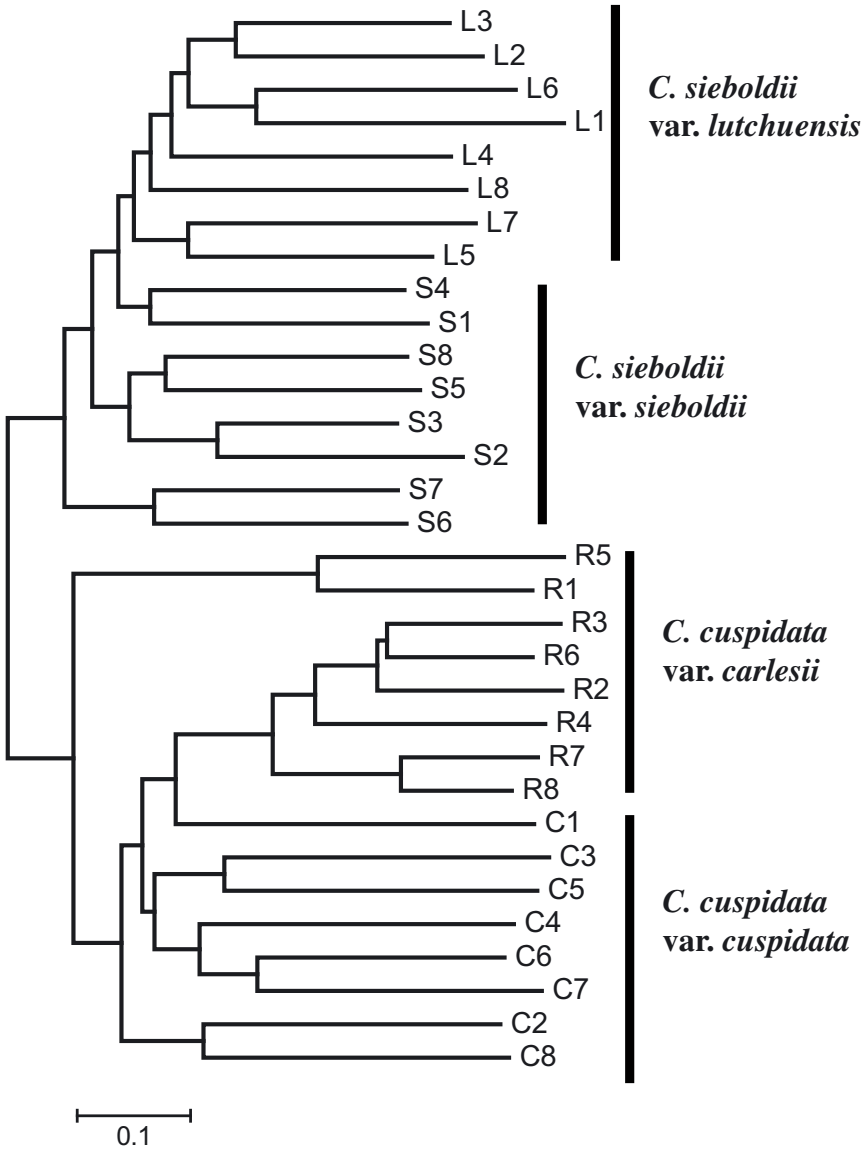

Figure 2. NJ dendrogram for individuals of *C. sieboldii* var. *lutchuensis* (L), *C. sieboldii* var. *sieboldii* (S), *C. cuspidata* var. *carlesii* (R) and *C. cuspidata* var. *cuspidata* (C) based on the proportion of shared alleles for 16 EST-SSR markers.

a metallothionein-like protein type 2 and the elongation factor 1-α were highly abundant transcripts. These transcripts found in poplar and *C. japonica* were also abundant in the material examined in the present study (Suppl. Tab. II[1]).

Although the number of ESTs in the present study is limited, the HMM search identified the ubiquitin protein family as the most abundant (Tab. I). Pavy et al. (2005) reported that the most abundant protein families in various tissue of *Picea glauca*, based on EST analysis, were the kinase domain, cytochrome P450, protein tyrosine kinase and the ubiquitin family. The observed differences may be the result of the characteristics of the inner bark. When the result was limited to putative unigenes relating to cell wall formation, the most abundant protein family was ubiquitin, followed by ubiquitin-conjugating enzyme and core histone H2A/H2B/H3/H4, with the number of potential unigenes 16, 9 and 9, respectively. Therefore, the putative unigenes with ubiquitin protein family was mostly (16 out of 18) related to cell wall formation. Ubiquitin is associated with protein degradation and cell death during the development of vascular tissue (Stephenson et al., 1996). The larger number of putative unigenes related to ubiquitin in the present library may relate to cell wall formation.

When our putative unigenes were compared with sequences relating to cell wall formation, 9.3% of the total number of transcripts exhibited similarities. The ratio in the present study was higher than or nearly the same as that reported for poplar (Sterky et al., 1998), *C. japonica* (Ujino-Ihara et al., 2000) and *Pinus taeda* (Allona et al., 1998), with ratios of 4%, 3% and ca. 10%, respectively. When we limited the MAIZEWALL database sequence to that listed in Table II of Sterky et al.

(1998) and adopted the same threshold (score > 100), 7.9% of our transcripts were considered to be related to cell wall formation.

4.2. Microsatellite mining and EST-SSR markers

We found a total of 286 (11.8%) putative unigenes with microsatellite motifs. The percentage of microsatellite-containing ESTs is reported to vary substantially among species, ranging from 2.65% for *Solanum tuberosum* to 16.82% for *Juglans regia*, with an average of 6.0% among dicotyledonous species (Kumpatla and Mukhopadhyay, 2005), and from 1.5% for maize to 4.7% for rice among cereal species (monocotyledons) (Kantety et al., 2002). Thus, unsurprisingly the abundance of microsatellites in our EST library is within the range reported for dicotyledons (although it should be noted that the thresholds for the number of microsatellite repeats differed between the two cited studies). Moreover, the abundance of SSRs, in terms of the microsatellite motifs, was highest for AG and AAG amongst di- and tri-SSRs, respectively (Tab. II). This trend is also found in most of the dicotyledonous species (Kumpatla and Mukhopadhyay, 2005). In coding regions, tri-SSRs were more frequent, while we found no di-SSRs (Tab. II), probably because selection against frameshift mutations removes di-SSR repeats from coding regions (Metzgar et al., 2000).

When we forcus on EST-SSRs in tree species, the percentage of microsatellite-containing ESTs for angiosperm species is generally high and AG motif is more rich, compared to gymnosperm species. For five *Eucalyptus* species, average frequency of microsatellite-containing EST was 12.9% with AG motif most common (Yasodha et al., 2008). *Quercus mongolica* var. *crispula* had 248 microsatellite-containing potential unigenes (11.6%) in all 2 140 potential unigenes, with AG motif most common (Ueno et al., 2008). *C. sieboldii* var. *sieboldii* in the present study (Tab. II) showd nealy the same trend as *Q. mongolica*. For *Pinus taeda* and *P. pinaster*, the percentage of SSR in their putative unigene was 1.2 and 2.1%, respectively (Chagné et al., 2004). The most common motif was AT and AG for *P. taeda* and *P. pinaster*, respectively. In *P. pinaster*, the occurrence of AT motof (47%) within di-SSR was nearly the same as that of AG motif (51%). *Picea* had only 183 contigs (0.93%) with at least one microsatellite in 20 275 putative unigenes (Rungis et al., 2004), while *Cryptomeria japonica* had 163 (3.6%) unique di- and tri- SSRs in about 4 500 cDNA clones (Moriguchi et al., 2003). For both *Picea* and *C. japonica*, AT motif was the most frequent.

Functional analysis of the putative unigenes and EST-SSRs based on GO slim terms (Fig. 1) revealed that the EST-SSRs were significantly more frequent than would be expected by chance in nucleotide-binding proteins (GO:0003676). Thirty-one putative unigenes with microsatellites were in this category; of these, 11 and 20 were estimated to be located in the coding and non-coding regions, respectively. More than half of the microsatellite-containing putative unigenes in GO:0003676 category had similarities to transcription and translation-related proteins. Complex regulation and machinery, including many protein-protein and/or protein-nucleic acid interactions, are needed in such processes and microsatellite or mono amino acid repeats may be important. Previous analyses of *Arabidopsis thaliana* and *Oryza sativa* proteins have shown that amino acid repeats are also overrepresented in their transcription factors (Zhang et al., 2006). The other suggested functions of the remaining microsatellite-containing putative unigenes included histone, DNA repair, and ribosomal proteins. Analysis of inner bark EST of *Quercus mongolica* (Ueno et al., 2008), produced similar results, with the GO:0003676 category being overrepresented.

In the present study, we developed 16 EST-SSR markers for *C. sieboldii* var. *sieboldii*, 12 of which have at least one di-SSR motif. The number of alleles per locus (*Na*) and observed heterozygosity (H_o) for these di-SSR markers ranged from 3 to 9, and from 0.2 to 0.9, respectively, in the *C. sieboldii* var. *sieboldii* individuals we examined (Tab. IV). Ueno et al. (2000 and 2003) surveyed the polymorphisms of 13 genomic anonymous di-SSR markers in 32 and 24 *C. sieboldii* var. *sieboldii* individuals, and found between six and 23 alleles per locus, with observed heterozygosity values ranging from 0.50 to 0.97. When allelic richness (Petit et al., 1998) for 10 diploid individuals is calculated for genomic di-SSR markers in the cited studies, it ranges from 4.6 to 12.5. The levels of variability still appear higher for anonymous genomic di-SSR markers than for EST-SSR markers with a di-SSR motif. This trend is to be expected, considering the more conservative nature of ESTs than that of anonymous genomic regions. However, in the systematic comparison made by Pashley et al. (2006), there were no significant differences between the levels of variability in transferable EST-SSR and gSSR markers. Moreover, the quality of the electropherograms for the EST-SSR markers was higher than that of the anonymous genomic SSRs (Pashley et al., 2006). In the present study, all of the EST-SSR markers had high quality electropherograms for samples from both of the two species and the two varieties. Similarly, 20 out of 44 candidate EST-SSR markers are reported to have been amplified successfully in samples from 23 *Picea* species (Rungis et al., 2004). Single nucleotide polymorphisms (SNPs) are frequently found within ESTs. However, analysis of SNPs markers for a large population is still costly. time-consuming and specific to a focal species, so it is inappropriate for practical purposes. EST-SSR markers are, in contrast, attractive tools for population genetic surveys with a wide scope for applying them to related species.

The cross-species PCR amplification over Fagaceae family in the present study (Tab. V) showed three markers (CcC02014, CcC02022 and CcC02069) were able to amplify all species tested. In order to develop common featurs for transferability, we focused on presence/absence of PCR products and primer locations on the putative unigenes. When both primers are located on the estimated coding region, they produced PCR product for all species tested. These markers were CcC01513, CcC02069, CcC02291 and CcC02375. The primer position may be an important factor to increase cross-species transferability, as in the case of *Helianthus* (Pashley et al., 2007). The cross-species transferability is also dependent on phylogenetic distance between species (Chagné et al., 2004).

This was also confirmed in the present study. *F. crenata* is the most distant species from *C. sieboldii* var. *sieboldii*, while *C. crenata* is the closest (Manos and Stanford, 2001). The transferability was the lowest for *F. crenata* and hightest for *C. crenata*.

Within *Castanopsis* genus, all of the markers in Table III gave clear electropherograms for all individuals in two species and two varieties. The discrimination of populations between *C. sieboldii* var. *sieboldii* and *C. cuspidata* var. *cuspidata* was clear (Fig. 2). Moreover, we have confirmed the usefulness of the markers in the present study through population analysis of *C. sieboldii* var. *sieboldii*, *C. sieboldii* var. *lutchuensis*, *C cuspidata* var. *cuspidata* and *C. cuspidata* var. *carlesii* (Aoki et al., unpublished). However, caution is required, when using putative single homologues of EST-SSR that have been amplified from *C. sieboldii* var. *sieboldii*, if the markers are applied to related species. Without sequencing, confirming the expected difference in allele size may be an alternative method to validate homologous amplification. Ten of the 16 EST-SSRs in Table IV had alleles with the expected difference in the repeat unit. The remaining loci probably experienced insertions or deletions in the vicinity of the microsatellites. We believe that the PCR primers designed in the present study will be of considerable value in future research into variations within *C. sieboldii* and related species.

Acknowledgements: The authors are grateful to H. Yoshimaru, Y. Tsuda, R. Kusano, K. Kimura, T. Kamijo, H. Setoguchi and K. Oono for collecting samples, to Y. Taguchi and Y. Komatsu for laboratory work and to T. Ujino-Ihara for valuable discussions. This research was supported by a grant for Research on Genetic Guidelines for Restoration Programs using Genetic Diversity Information from the Ministry of Environment, Japan.

REFERENCES

Allona I., Quinn M., Shoop E., Swope K., St Cyr S., Carlis J., Riedl J., Retzel E., Campbell M.M., Sederoff R., and Whetten R.W., 1998. Analysis of xylem formation in pine by cDNA sequencing. Proc. Natl. Acad. Sci. USA 95: 9693–9698.

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Apweiler R., Bairoch A., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N., and Yeh L.S., 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 32: D115–D119.

Bhalerao R., Keskitalo J., Sterky F., Erlandsson R., Bjorkbacka H., Birve S.J., Karlsson J., Gardestrom P., Gustafsson P., Lundeberg J., and Jansson S., 2003. Gene expression in autumn leaves. Plant Physiol. 131: 430–442.

Bommer U.A. and Thiele B.J., 2004. The translationally controlled tumour protein (TCTP). Int. J. Biochem. Cell Biol. 36: 379–385.

Cans C., Passer B.J., Shalak V., Nancy-Portebois V., Crible V., Amzallag N., Allanic D., Tufino R., Argentini M., Moras D., Fiucci G., Goud B., Mirande M., Amson R., and Telerman A., 2003. Translationally controlled tumor protein acts as a guanine nucleotide dissociation inhibitor on the translation elongation factor eEF1A. Proc. Natl. Acad. Sci. USA 100: 13892–13897.

Chagne D., Chaumeil P., Ramboer A., Collada C., Guevara A., Cervera M.T., Vendramin G.G., Garcia V., Frigerio J.M., Echt C., Richardson T., and Plomion C., 2004. Cross-species transferability and mapping of genomic and cDNA SSRs in pines. Theor. Appl. Genet. 109: 1204–1214.

Chang S., Puryear J., and Cairney J., 1993. A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. 11: 113–116.

Cho C., Lee H., Chung E., Kim K., Heo J., Kim J., Chung J., Ma Y., Fukui K., Lee D., Kim D., Chung Y., and Lee J., 2007. Molecular characterization of the soybean L-asparaginase gene induced by low temperature stress. Mol. Cells 23: 280–286.

Delectis florae reipublicae popularis sinicae agendae academiae sinicae edita, 1998. Flora : reipublicae popularis sinicae (in Chinese) Science Press, Beijing, Vol. 22, 66–67.

Dieringer D. and Schlotterer C., 2003. Microsatellite analyzer (MSA): a platform independent analysis tool for large microsatellite data sets. Mol. Ecol. Notes 3: 167–169.

Eveno E., Collada C., Guevara M.A., Leger V., Soto A., Diaz L., Leger P., Gonzalez-Martinez S.C., Cervera M.T., Plomion C., and Garnier-Gere P.H., 2008. Contrasting patterns of selection at *Pinus pinaster* Ait. Drought stress candidate genes as revealed by genetic differentiation analyses. Mol. Biol. Evol. 25: 417–437.

Ewing B. and Green P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Ewing B., Hillier L., Wendl M.C., and Green P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175–185.

Ewing R.M., Ben Kahla A., Poirot O., Lopez F., Audic S., and Claverie J.M. (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res. 9: 950–959

Finn R.D., Tate J., Mistry J., Coggill P.C., Sammut S.J., Hotz H.R., Ceric G., Forslund K., Eddy S.R., Sonnhammer E.L., and Bateman A., 2008. The Pfam protein families database. Nucleic Acids Res. 36: D281–D288.

Fukuoka H., Nunome T., Minamiyama Y., Kono I., Namiki N., and Kojima A., 2005. Read2Marker: a data processing tool for microsatellite marker development from a large data set. Biotechniques 39: 472, 474, 476.

Girke T., Lauricha J., Tran H., Keegstra K., and Raikhel N., 2004. The cell wall navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. Plant Physiol. 136: 3003–3008; discussion 3001.

Grant M. and Bevan M.W., 1994. Asparaginase gene expression is regulated in a complex spatial and temporal pattern in nitrogen-sink tissues. Plant J. 5: 695–704.

Green P., Documentation for phrap and cross_match. 1999. [online] Available from http://bozeman.mbt.washington.edu/phrap.docs/phrap.html [accessed 7 March 2007].

Guillaumie S., San-Clemente H., Deswarte C., Martinez Y., Lapierre C., Murigneux A., Barriere Y., Pichon M., and Goffner D., 2007. MAIZEWALL. Database and developmental gene expression profiling of cell wall biosynthesis and assembly in maize. Plant Physiol. 143: 339–363.

Iseli C., Jongeneel C.V., and Bucher P., 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 138–148.

Kado T., Yoshimaru H., Tsumura Y., and Tachida H., 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). Genetics 164: 1547–1559.

Kane N.C. and Rieseberg L.H., 2007. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. Genetics 175: 1823–1834.

Kantety R.V., La Rota M., Matthews D.E., and Sorrells M.E., 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol. Biol. 48: 501–510.

Kobayashi S. and Hiroki S., 2003. Patterns of occurrence of hybrids of *Castanopsis cuspidata* and *C. sieboldii* in the IBP Minamata Special Research Area, Kumamoto Prefecture, Japan. J. Phytogeogr. Taxon. 51: 63–67.

Kobayashi Y. and Sugawa T., 1959. Identification of wood of some *Castanopsis* species in Japan (in Japanese with English abstract). Bull. Gov. For. Exp. Stn. 118: 139–178.

Kondo H., Tahira T., Hayashi H., Oshima K., and Hayashi K., 2000. Microsatellite genotyping of post-PCR fluorescently labeled markers. Biotechniques 29: 868–872.

Kumpatla S.P. and Mukhopadhyay S., 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48: 985–998.

Manos P.S. and Stanford A.M., 2001. The historical biogeography of Fagaceae: Tracking the tertiary history of temperate and subtropical forests of the Northern Hemisphere. Int. J. Plant Sci. 162: S77–S93.

Metzgar D., Bytof J., and Wills C., 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 10: 72–80.

Moriguchi Y., Iwata H., Ujino-Ihara T., Yoshimura K., Taira H., and Tsumura Y., 2003. Development and characterization of microsatellite markers for *Cryptomeria japonica* D. Don. Theor. Appl. Genet. 106: 751–758.

Murray M.G. and Thompson W.F., 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8: 4321–4325.

Nanjo T., Futamura N., Nishiguchi M., Igasaki T., Shinozaki K., and Shinohara K., 2004. Characterization of full-length enriched expressed sequence tags of stress-treated poplar leaves. Plant Cell Physiol. 45: 1738–1748.

Navabpour S., Morris K., Allen R., Harrison E., S A.H.-M., and Buchanan-Wollaston V., 2003. Expression of senescence-enhanced genes in response to oxidative stress. J. Exp. Bot. 54: 2285–2292.

Nei M. and Kumar S., 2000. Molecular evolution and phylogenetics, Oxford University Press, New York, 333 p.

Parkinson J., Anthony A., Wasmuth J., Schmid R., Hedley A., and Blaxter M., 2004. PartiGene–constructing partial genomes. Bioinformatics 20: 1398–1404.

Parkinson J., Guiliano D.B., and Blaxter M., 2002. Making sense of EST sequences by CLOBBing them. BMC Bioinformatics 3: 31.

Pashley C.H., Ellis J.R., McCauley D.E., and Burke J.M., 2006. EST databases as a source for molecular markers: lessons from *Helianthus*. J. Hered. 97: 381–388.

Pavy N., Paule C., Parsons L., Crow J.A., Morency M.J., Cooke J., Johnson J.E., Noumen E., Guillet-Claude C., Butterfield Y., Barber S., Yang G., Liu J., Stott J., Kirkpatrick R., Siddiqui A., Holt R., Marra M., Seguin A., Retzel E., Bousquet J., and MacKay J., 2005. Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. BMC Genomics 6: 144.

Petit R.J., El Mousadik A., and Pons O., 1998. Identifying populations for conservation on the basis of genetic markers. Conserv. Biol. 12: 844–855.

Raes J., Rohde A., Christensen J.H., Van de Peer Y., and Boerjan W., 2003. Genome-wide characterization of the lignification toolbox in *Arabidopsis*. Plant Physiol. 133: 1051–1071.

Rozen S. and Skaletsky H.J., 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S.A. and Misener S. (Eds.), Bioinformatics methods and protocols: Methods in molecular biology, Humana Press, Totowa, pp. 365–386.

Rungis D., Berube Y., Zhang J., Ralph S., Ritland C.E., Ellis B.E., Douglas C., Bohlmann J., and Ritland K., 2004. Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. Theor. Appl. Genet. 109: 1283–1294.

Stephenson P., Collins B.A., Reid P.D., and Rubinstein B., 1996. Localization of ubiquitin to differentiating vascular tissues. Am. J. Bot. 83: 140–147.

Sterky F., Regan S., Karlsson J., Hertzberg M., Rohde A., Holmberg A., Amini B., Bhalerao R., Larsson M., Villarroel R., Van Montagu M., Sandberg G., Olsson O., Teeri T.T., Boerjan W., Gustafsson P., Uhlen M., Sundberg B., and Lundeberg J., 1998. Gene discovery in the wood-forming tissues of poplar: analysis of 5, 692 expressed sequence tags. Proc. Natl. Acad. Sci. USA 95: 13330–13335.

Sterky F., Bhalerao R.R., Unneberg P., Segerman B., Nilsson P., Brunner A.M., Charbonnel-Campaa L., Lindvall J.J., Tandre K., Strauss S.H., Sundberg B., Gustafsson P., Uhlen M., Bhalerao R.P., Nilsson O., Sandberg G., Karlsson J., Lundeberg J., and Jansson S., 2004. A Populus EST resource for plant functional genomics. Proc. Natl. Acad. Sci. USA 101: 13951–13956.

Tamura K., Dudley J., Nei M., and Kumar S., 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24: 1596–1599.

Temnykh S., DeClerck G., Lukashova A., Lipovich L., Cartinhour S., and McCouch S., 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11: 1441–1452.

Tsumura Y., Kado T., Takahashi T., Tani N., Ujino-Ihara T., and Iwata H., 2007. Genome scan to detect genetic structure and adaptive genes of natural populations of *Cryptomeria japonica*. Genetics 176: 2393–2403.

Ueno S., Taguchi Y., and Tsumura Y., 2008. Microsatellite markers derived from *Quercus mongolica* var. *crispula* (Fagaceae) inner bark expressed sequence tags. Genes Genet. Syst. 83: 179–187.

Ueno S., Yoshimaru H., Kawahara T., and Yamamoto S., 2000. Isolation of microsatellite markers in *Castanopsis cuspidata* var. *sieboldii* Nakai from an enriched library. Mol. Ecol. 9: 1188–1190.

Ueno S., Yoshimaru H., Kawahara T., and Yamamoto S., 2003. A further six microsatellite markers for *Castanopsis cuspidata* var. *sieboldii* Nakai. Conserv. Genet. 4: 813–815.

Ujino-Ihara T., Yoshimura K., Ugawa Y., Yoshimaru H., Nagasaka K., and Tsumura Y., 2000. Expression analysis of ESTs derived from the inner bark of *Cryptomeria japonica*. Plant Mol. Biol. 43: 451–457.

Yamada H. and Miyaura T., 2003. Geographic occurrence of intermediate type between *Castanopsis sieboldii* and *C. cuspidata* (Fagaceae) based on the structure of leaf epidermis. J. Plant Res. 116: 477–482.

Yamaguchi-Shinozaki K. and Shinozaki K., 1993. The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of rd22, a gene responsive to dehydration stress in *Arabidopsis thaliana*. Mol. Gen. Genet. 238: 17–25.

Yamanaka T., 1966. Problems of *Castanopsis cuspidata* Schottky (in Japanese with English abstract). Bull. Fac. Educ., Kochi Univ. 18: 65–73.

Yamazaki T. and Mashiba S., 1987a. A taxonomical revision of *Castanopsis cuspidata* (Thunb.) Schottky and the allies in Japan, Korea and Taiwan (1). J. Jap. Bot. 62: 289–298.

Yamazaki T. and Mashiba S., 1987b. A taxonomical revision of *Castanopsis cuspidata* (Thunb.) Schottky and the allies in Japan, Korea and Taiwan (2). J. Jap. Bot. 62: 332–339.

Yasodha R., Sumathi R., Chezhian P., Kavitha S., and Ghosh M., 2008. *Eucalyptus* microsatellites mined in silico: survey and evaluation. J. Genet. 87: 21–25.

Zhang L., Yu S., Cao Y., Wang J., Zuo K., Qin J., and Tang K., 2006. Distributional gradient of amino acid repeats in plant proteins. Genome 49: 900–905.

Ann. For. Sci. 66 (2009) 509

Online Material

Supplementary Table I. Sampling sites for polymorphism screening.

Species	Sample locality	Longitude (°)	Latitude (°)
C. sieboldii var. *sieboldii*	Ibaraki	140.52	36.02
	Tokyo	139.82	33.11
	Toyama	137.01	36.89
	Aichi	137.30	35.09
	Wakayama	135.61	33.52
	Shimane	133.01	35.51
	Kochi	133.15	33.15
	Nagasaki	129.20	34.16
	Miyazaki	131.42	32.17
	Kagoshima	130.63	30.35
C. sieboldii var. *lutchuensis*	Okinawa	127.97	26.67
C. cuspidata var. *cuspidata*	Mie	136.63	34.46
	Hiroshima	132.21	34.25
	Kagawa	134.37	34.21
	Kumamoto	130.74	32.83
C. cuspidata var. *carlesii*	Taiwan	121.73	25.02

Supplementary Table II. The characteristics of abundant clusters (cluster size > 5) in the *C. sieboldii* inner bark cDNA library.

Cluster ID	Contig No.	Cluster Size	Nr blastx top hit			Pfam top hit	
			Accession No.	Description [species]	E-value	Description	E-value
CcC00008	1	39	gb|AAV36561.1|	RD22-like protein [*Vitis vinifera*]	8E-45	BURP domain	2.3E-51
CcC00064	1	19	emb|CAC39481.2|	Metallothionein-like protein [*Quercus suber*]	7E-25	Metallothionein	2.7E-46
CcC00021	1	18	gb|AAM23265.1|	L-asparaginase [*Glycine max*]	1E-159	Asparaginase	1.8E-119
CcC00068	1	17	ref|NP_566536.1|	Asparaginase [*Arabidopsis thaliana*]	1E-58	–	–
CcC00033	1	16	ref|NP_194443.1|	Translation initiation factor [*Arabidopsis thaliana*]	7E-56	Translation initiation factor SUI1	2.7E-43
CcC00183	1	15	ref|NP_197617.1|	Ribonuclease/transcriptional repressor [*Arabidopsis thaliana*]	1E-111	CAF1 family ribonuclease	1E-117
CcC00497	1	15	dbj|BAC79443.1|	Catalase [*Acacia ampliceps*]	1E-176	Catalase	1.4E-275
CcC00353	1	13	gb|AAD10032.1|	Translationally controlled tumor protein [*Hevea brasiliensis*]	4E-77	Translationally controlled tumor protein	2.6E-105
CcC00583	1	13	ref|NP_187955.2|	Unknown protein [*Arabidopsis thaliana*]	1E-24	–	–
CcC00038	1	12	ref|NP_190815.1|	Unknown protein [*Arabidopsis thaliana*]	1E-38	Harpin-induced protein 1 (Hin1)	5.5E-24
CcC00304	1	12	gb|AAK69513.1|	Putative transcription factor [*Vitis vinifera*]	2E-07	ABA/WDS induced protein	1.1E-47
CcC00506	1	12	emb|CAC09928.1|	Hypothetical protein [*Catharanthus roseus*]	9E-17	–	–
CcC00126	1	10	–	–	–	–	–
CcC00216	1	10	gb|ABB02647.1|	Unknown [*Solanum tuberosum*]	5E-68	Ribosomal protein S8	4.3E-62
CcC00071	1	9	gb|AAA98491.1|	Anionic peroxidase [*Petroselinum crispum*]	1E-126	Peroxidase	2E-128
CcC00214	1	9	gb|AAD26942.1|	Zinc-finger protein 1 [*Datisca glomerata*]	3E-50	–	–
CcC00524	1	9	gb|AAN77145.1|	Fiber protein Fb2 [*Gossypium barbadense*]	1E-72	Drought induced 19 protein (Di19)	1.7E-93
CcC00231	1	8	dbj|BAC23049.1|	Elongation factor 1-α [*Solanum tuberosum*]	1E-114	Elongation factor Tu GTP binding domain	1.5E-90
CcC00231	2	8	gb|ABA12218.1|	Translation elongation factor 1A-2 [*Gossypium hirsutum*]	0	Elongation factor Tu GTP binding domain	1.5E-90
CcC00285	1	8	gb|AAV66332.1|	Ethylene response factor 3 [*Cucumis sativus*]	1E-35	AP2 domain	5.8E-21
CcC00345	1	8	gb|AAU04436.1|	MAPKK [*Lycopersicon esculentum*]	1E-134	Protein kinase domain	9E-78
CcC00473	1	8	ref|NP_190848.1|	DNA binding/zinc ion binding [*Arabidopsis thaliana*]	4E-25	AN1-like Zinc finger	2.3E-19
CcC00516	1	8	ref|NP_199299.1|	Unknown protein [*Arabidopsis thaliana*]	1E-10	–	–
CcC00653	1	8	gb|AAG40371.1|	AT4g27960 [*Arabidopsis thaliana*]	3E-77	Ubiquitin-conjugating enzyme	3.3E-77
CcC00829	1	8	gb|AAB95222.1|	Cytosolic ascorbate peroxidase [*Fragaria* × *ananassa*]	1E-116	Peroxidase	3.7E-66
CcC00914	1	8	emb|CAB85630.1|	Putative metallothionein-like protein [*Vitis vinifera*]	9E-13	–	–
CcC00948	1	8	–	–	–	–	–
CcC00053	1	7	–	–	–	–	–
CcC00091	1	7	emb|CAC84116.1|	Peptidylprolyl isomerase (cyclophilin) [*Betula pendula*]	1E-81	Cyclophilin type peptidyl-prolyl cis-tr	4.4E-125
CcC00121	1	7	gb|AAK29409.1|	S-adenosyl-L-methionine synthetase [*Elaeagnus umbellata*]	1E-107	S-adenosylmethionine synthetase, cent	1.1E-38
CcC00207	1	7	sp|Q40161|GP1_LYCES	Polygalacturonase isoenzyme 1 beta subunit [*Solanum lycopersicum*]	2E-52	BURP domain	1.5E-88
CcC00234	1	7	ref|NP_180826.2|	Putative synaptobrevin protein [*Arabidopsis thaliana*]	2E-86	–	–
CcC00352	1	7	ref|NP_187955.2|	Unknown protein [*Arabidopsis thaliana*]	4E-65	YT521-B-like family	1.8E-60
CcC00651	1	7	ref|NP_197927.1|	Ubiquinol-cytochrome-c reductase [*Arabidopsis thaliana*]	2E-50	Ubiquinol-cytochrome C reductase complex 14k	1.6E-39
CcC00020	1	6	gb|AAL24250.1|	At1g19180/T29M8_5 [*Arabidopsis thaliana*]	2E-15	–	–
CcC00052	1	6	gb|AAT68207.1|	Unknown [*Cynodon dactylon*]	4E-12	Protein of unknown function (DUF581)	2.3E-19
CcC00061	1	6	dbj|BAD29340.1|	Hypothetical protein [*Oryza sativa*]	1E-12	–	–
CcC00063	1	6	gb|AAM65077.1|	Unknown [*Arabidopsis thaliana*]	5E-18	–	–
CcC00186	1	6	gb|AAF79874.1|	T7N9.16 [*Arabidopsis thaliana*]	5E-51	Protein of unknown function (DUF569)	9.5E-87
CcC00208	1	6	dbj|BAA33810.1|	Phi-1 [*Nicotiana tabacum*]	1E-131	Phosphate-induced protein 1 conserved region	5.9E-191
CcC00235	1	6	gb|ABB17004.1|	Ribosomal protein S7-like protein [*Solanum tuberosum*]	2E-87	Ribosomal protein S7e	7.1E-88
CcC00256	1	6	gb|AAV51937.1|	AP2/EREBP transcription factor ERF-2 [*Gossypium hirsutum*]	4E-33	AP2 domain	2.4E-21
CcC00426	1	6	gb|AAF27931.1|	14-3-3-like protein [*Euphorbia esula*]	1E-110	14-3-3 protein	6.3E-147
CcC00873	1	6	dbj|BAA10929.1|	Cytochrome P450 like_TBP [*Nicotiana tabacum*]	5E-53	–	–

–: No hits found.

Supplementary Table III. Distance matrix among cDNA libraries. CcC: *C. sieboldii* in the present study, Leple M: mild draught stress-treated *Populus tremula* × *Populus alba* leaves, Leple S: sevear draught stress-treated *Populus tremula* x *Populus alba* leaves, Nanjo: mixture of several stress-treated *Populus nigra* leaves. Description for other libraries can be found at http://www.populus.db.umu.se/.

Library	CcC	A+B	C	F	G	I	K	L	M	N	P	Q	R	S	T	U	UB	V	X	Leple M	Leple S	Nanjo
CcC	0.00	3.29	3.66	3.52	3.59	2.71	3.78	3.10	3.63	3.92	3.81	3.23	3.75	3.55	3.73	3.80	3.79	3.26	2.69	2.60	2.38	3.25
A+B	3.29	0.00	0.78	0.68	1.01	1.62	0.91	0.76	0.76	0.86	0.88	1.16	0.86	0.91	0.77	0.73	0.85	0.85	1.45	1.61	2.18	1.03
C	3.66	0.78	0.00	1.04	1.41	1.55	1.17	1.08	1.11	0.96	1.18	1.04	1.12	0.29	0.32	0.66	1.11	1.35	1.81	1.50	2.15	1.49
F	3.52	0.68	1.04	0.00	0.81	1.96	0.76	0.97	0.56	0.65	0.48	1.30	0.71	1.15	0.93	0.67	0.73	0.72	1.40	1.90	2.47	0.92
G	3.59	1.01	1.41	0.81	0.00	2.41	0.36	1.24	0.97	0.75	0.68	1.89	0.53	1.52	1.24	0.88	0.44	0.62	1.94	2.26	2.89	0.45
I	2.71	1.62	1.55	1.96	2.41	0.00	2.34	1.44	1.96	2.21	2.25	0.93	2.22	1.38	1.71	2.02	2.29	2.07	1.45	0.85	0.92	2.25
K	3.78	0.91	1.17	0.76	0.36	2.34	0.00	1.17	0.84	0.45	0.53	1.77	0.29	1.31	0.98	0.60	0.15	0.74	2.00	2.24	2.88	0.73
L	3.10	0.76	1.08	0.97	1.24	1.44	1.17	0.00	0.78	1.11	1.06	1.04	1.01	1.12	1.12	1.11	1.11	1.05	1.30	1.64	2.14	1.21
M	3.63	0.76	1.11	0.56	0.97	1.96	0.84	0.78	0.00	0.61	0.45	1.24	0.63	1.22	1.02	0.80	0.79	0.87	1.44	2.09	2.59	1.15
N	3.92	0.86	0.96	0.65	0.75	2.21	0.45	1.11	0.61	0.00	0.37	1.52	0.34	1.11	0.78	0.41	0.41	0.91	1.89	2.19	2.81	1.06
P	3.81	0.88	1.18	0.48	0.68	2.25	0.53	1.06	0.45	0.37	0.00	1.55	0.41	1.32	1.03	0.66	0.50	0.81	1.71	2.25	2.83	0.97
Q	3.23	1.16	1.04	1.30	1.89	0.93	1.77	1.04	1.24	1.52	1.55	0.00	1.60	0.92	1.15	1.40	1.72	1.59	1.09	1.25	1.58	1.85
R	3.75	0.86	1.12	0.71	0.53	2.22	0.29	1.01	0.63	0.34	0.41	1.60	0.00	1.25	0.94	0.62	0.29	0.71	1.86	2.20	2.81	0.83
S	3.55	0.91	0.29	1.15	1.52	1.38	1.31	1.12	1.22	1.11	1.32	0.92	1.25	0.00	0.42	0.83	1.27	1.39	1.74	1.32	1.95	1.56
T	3.73	0.77	0.32	0.93	1.24	1.71	0.98	1.12	1.02	0.78	1.03	1.15	0.94	0.42	0.00	0.48	0.94	1.17	1.84	1.64	2.28	1.36
U	3.80	0.73	0.66	0.67	0.88	2.02	0.60	1.11	0.80	0.41	0.66	1.40	0.62	0.83	0.48	0.00	0.56	0.97	1.87	1.92	2.57	1.10
UB	3.79	0.85	1.11	0.73	0.44	2.29	0.15	1.11	0.79	0.41	0.50	1.72	0.29	1.27	0.94	0.56	0.00	0.80	1.98	2.22	2.87	0.79
V	3.26	0.85	1.35	0.72	0.62	2.07	0.74	1.05	0.87	0.91	0.81	1.59	0.71	1.39	1.17	0.97	0.80	0.00	1.56	1.96	2.51	0.61
X	2.69	1.45	1.81	1.40	1.94	1.45	2.00	1.30	1.44	1.89	1.71	1.09	1.86	1.74	1.84	1.87	1.98	1.56	0.00	1.61	1.74	1.77
Leple M	2.60	1.61	1.50	1.90	2.26	0.85	2.24	1.64	2.09	2.19	2.25	1.25	2.20	1.32	1.64	1.92	2.22	1.96	1.61	0.00	0.88	2.04
Leple S	2.38	2.18	2.15	2.47	2.89	0.92	2.88	2.14	2.59	2.81	2.83	1.58	2.81	1.95	2.28	2.57	2.87	2.51	1.74	0.88	0.00	2.65
Nanjo	3.25	1.03	1.49	0.92	0.45	2.25	0.73	1.21	1.15	1.06	0.97	1.85	0.83	1.56	1.36	1.10	0.79	0.61	1.77	2.04	2.65	0.00

Supplementary Table IV. SSRs found in potential unigenes by SSRIT.pl script and estimation of SSR location.

Potential unigene ID	SSR number	Motif length	Motif	Number of motif	Stat position	End position	SSR location
CcC00006_1	1	2	at	9	69	86	Non-coding
CcC00019_1	1	3	gaa	6	35	52	Coding
CcC00021_1	1	2	ga	10	34	53	Non-coding
CcC00024_1	1	2	ct	16	12	43	Non-coding
CcC00033_1	1	4	gagc	5	424	443	Non-coding
CcC00034_1	1	2	tc	15	21	50	Unknown
CcC00037_1	1	2	tc	17	54	87	Non-coding
CcC00048_1	1	2	tc	9	86	103	Unknown
CcC00052_1	1	2	ct	15	17	46	Non-coding
CcC00053_1	1	2	ga	40	9	88	Non-coding
CcC00055_1	1	2	at	10	92	111	Unknown
CcC00059_1	1	3	aac	6	193	210	Coding
CcC00063_1	1	3	aga	6	373	390	Coding
CcC00068_1	1	2	ga	13	33	58	Non-coding
CcC00072_1	1	2	ag	9	14	31	Non-coding
CcC00083_1	1	2	tc	11	96	117	Non-coding
CcC00087_1	1	2	tc	13	32	57	Non-coding
CcC00110_1	1	3	tat	6	541	558	Unknown
CcC00112_1	1	2	tc	20	38	77	Non-coding
CcC00114_1	1	3	tta	6	59	76	Non-coding
CcC00118_1	1	3	ccg	6	328	345	Coding
CcC00127_1	1	2	ct	9	11	28	Non-coding
CcC00132_1	1	2	at	12	377	400	Non-coding
CcC00134_1	1	2	at	12	473	496	Non-coding
CcC00141_1	1	3	aca	6	21	38	Coding
CcC00141_1	2	3	aca	6	202	219	Coding
CcC00143_1	1	2	ta	16	563	594	Non-coding
CcC00146_1	1	2	ag	12	128	151	Non-coding
CcC00150_1	1	2	ct	21	14	55	Non-coding
CcC00152_1	1	3	tct	7	239	259	Coding
CcC00155_1	1	3	tct	8	119	142	Coding
CcC00157_1	1	2	ta	12	365	388	Non-coding
CcC00158_1	1	2	ag	9	105	122	Non-coding
CcC00171_1	1	2	ct	13	170	195	Non-coding
CcC00172_1	1	2	ta	11	124	145	Non-coding
CcC00176_1	1	3	gga	6	304	321	Coding
CcC00206_1	1	3	ttc	6	397	414	Unknown
CcC00231_2	1	2	ct	9	8	25	Non-coding
CcC00233_1	1	3	ttc	6	17	34	Non-coding
CcC00248_1	1	2	tc	10	9	28	Non-coding
CcC00248_1	2	2	ct	11	109	130	Non-coding
CcC00253_1	1	2	tc	14	27	54	Non-coding
CcC00260_1	1	3	gga	7	276	296	Coding
CcC00270_1	1	3	ctt	6	101	118	Coding
CcC00296_1	1	3	tta	7	496	516	Non-coding
CcC00301_1	1	2	ct	9	25	42	Non-coding
CcC00307_1	1	2	ct	13	63	88	Non-coding
CcC00314_1	1	2	ac	16	76	107	Unknown
CcC00314_1	2	2	at	11	108	129	Unknown
CcC00315_1	1	2	ct	10	63	82	Non-coding
CcC00321_1	1	3	tct	6	53	70	Non-coding
CcC00329_1	1	2	tc	10	25	44	Non-coding
CcC00338_1	1	2	ag	12	78	101	Non-coding
CcC00338_1	2	3	aca	6	360	377	Coding
CcC00338_1	3	3	gga	8	392	415	Coding
CcC00348_1	1	3	caa	6	148	165	Coding
CcC00370_1	1	3	aag	6	163	180	Coding
CcC00370_1	2	3	aag	8	259	282	Coding

Supplementary Table IV. Continued.

Potential unigene ID	SSR number	Motif length	Motif	Number of motif	Stat position	End position	SSR location
CcC00370_1	3	3	aag	6	292	309	Coding
CcC00374_1	1	2	tc	10	68	87	Unknown
CcC00376_1	1	2	ag	12	16	39	Non-coding
CcC00385_1	1	2	at	12	541	564	Non-coding
CcC00397_1	1	2	ct	15	74	103	Non-coding
CcC00400_1	1	3	gaa	8	295	318	Coding
CcC00418_1	1	3	caa	6	319	336	Coding
CcC00428_1	1	2	ct	9	61	78	Non-coding
CcC00435_1	1	2	ca	14	9	36	Unknown
CcC00440_1	1	3	cca	8	175	198	Coding
CcC00446_1	1	2	ag	13	125	150	Non-coding
CcC00454_1	1	4	gtca	56	9	232	Unknown
CcC00473_1	1	3	ctc	6	285	302	Coding
CcC00487_1	1	3	gta	7	14	34	Coding
CcC00495_1	1	3	tct	7	6	26	Non-coding
CcC00511_1	1	2	tc	10	17	36	Non-coding
CcC00524_1	1	2	ct	9	78	95	Non-coding
CcC00534_1	1	3	gga	6	312	329	Coding
CcC00535_1	1	2	ta	9	174	191	Non-coding
CcC00540_1	1	3	tct	6	99	116	Coding
CcC00556_1	1	3	tct	14	85	126	Non-coding
CcC00560_1	1	3	atc	9	405	431	Coding
CcC00567_1	1	2	ct	19	49	86	Non-coding
CcC00583_1	1	2	tc	11	218	239	Non-coding
CcC00599_1	1	2	ct	10	140	159	Non-coding
CcC00609_1	1	3	gga	6	151	168	Coding
CcC00610_1	1	2	ta	12	334	357	Non-coding
CcC00634_1	1	2	tc	11	45	66	Unknown
CcC00637_1	1	2	ct	12	27	50	Non-coding
CcC00646_1	1	2	ta	11	18	39	Non-coding
CcC00651_1	1	2	ga	13	29	54	Non-coding
CcC00660_1	1	2	ct	9	90	107	Non-coding
CcC00660_1	2	3	aac	6	333	350	Coding
CcC00678_1	1	2	ag	14	25	52	Non-coding
CcC00678_1	2	2	ga	11	79	100	Non-coding
CcC00705_1	1	3	tcc	6	31	48	Coding
CcC00729_1	1	2	tc	21	42	83	Unknown
CcC00729_1	2	2	ag	16	159	190	Unknown
CcC00735_1	1	2	ag	11	13	34	Non-coding
CcC00744_1	1	4	ttat	5	79	98	Unknown
CcC00749_1	1	2	tc	17	156	189	Non-coding
CcC00750_1	1	2	tc	14	128	155	Unknown
CcC00760_1	1	2	ct	13	9	34	Non-coding
CcC00761_1	1	3	aca	7	316	336	Coding
CcC00761_1	2	3	agg	6	463	480	Coding
CcC00764_1	1	3	ttc	8	164	187	Non-coding
CcC00767_1	1	3	cac	8	95	118	Coding
CcC00769_1	1	2	tc	10	63	82	Non-coding
CcC00782_1	1	3	gaa	6	361	378	Coding
CcC00785_1	1	2	ct	15	12	41	Non-coding
CcC00795_1	1	3	ttc	19	288	344	Non-coding
CcC00804_1	1	2	ct	11	39	60	Non-coding
CcC00816_1	1	3	gaa	6	170	187	Coding
CcC00816_1	2	3	gaa	6	276	293	Coding
CcC00817_1	1	4	tctt	6	27	50	Non-coding
CcC00824_1	1	2	ct	11	39	60	Unknown
CcC00824_1	2	2	ct	14	206	233	Unknown
CcC00845_1	1	2	ga	11	555	576	Unknown

Supplementary Table IV. Continued.

Potential unigene ID	SSR number	Motif length	Motif	Number of motif	Stat position	End position	SSR location
CcC00845_1	2	3	tgt	6	197	214	Unknown
CcC00849_1	1	3	caa	6	233	250	Coding
CcC00855_1	1	3	gga	6	315	332	Coding
CcC00857_1	1	2	tc	13	22	47	Non-coding
CcC00857_1	2	2	ca	17	47	80	Non-coding
CcC00866_1	1	2	ct	10	45	64	Non-coding
CcC00878_1	1	3	tct	7	42	62	Coding
CcC00879_1	1	4	acca	5	274	293	Non-coding
CcC00884_1	1	2	gt	10	385	404	Unknown
CcC00897_1	1	2	tc	12	11	34	Non-coding
CcC00936_1	1	3	aat	7	30	50	Non-coding
CcC00948_1	1	2	ct	16	25	56	Non-coding
CcC00954_1	1	2	tc	9	100	117	Non-coding
CcC00959_1	1	3	cca	6	354	371	Coding
CcC00966_1	1	3	gaa	10	55	84	Coding
CcC00968_1	1	2	tc	11	698	719	Non-coding
CcC00970_1	1	3	gag	6	315	332	Non-coding
CcC00972_1	1	2	ct	55	10	119	Non-coding
CcC00972_2	1	2	ga	55	8	117	Non-coding
CcC00973_1	1	2	tc	14	116	143	Non-coding
CcC00983_1	1	2	ct	71	9	150	Unknown
CcC00989_1	1	2	ct	14	37	64	Non-coding
CcC00997_1	1	3	gga	6	315	332	Coding
CcC01013_1	1	3	acc	7	241	261	Coding
CcC01013_1	2	3	cca	6	335	352	Coding
CcC01040_1	1	2	ag	18	132	167	Non-coding
CcC01059_1	1	3	gct	6	399	416	Coding
CcC01065_1	1	2	ag	11	494	515	Non-coding
CcC01067_1	1	3	agg	6	34	51	Coding
CcC01073_1	1	2	ga	13	160	185	Non-coding
CcC01078_1	1	3	cac	6	123	140	Coding
CcC01090_1	1	2	tc	9	13	30	Non-coding
CcC01097_1	1	2	at	12	510	533	Non-coding
CcC01111_1	1	3	gaa	11	91	123	Non-coding
CcC01135_1	1	3	ctt	8	166	189	Coding
CcC01146_1	1	2	tc	9	20	37	Non-coding
CcC01157_1	1	2	ag	9	16	33	Non-coding
CcC01160_1	1	3	tcc	7	141	161	Coding
CcC01167_1	1	3	gga	6	320	337	Coding
CcC01175_1	1	2	ct	9	19	36	Non-coding
CcC01183_1	1	2	ct	9	26	43	Non-coding
CcC01191_1	1	3	acc	6	85	102	Coding
CcC01225_1	1	2	ta	9	603	620	Non-coding
CcC01226_1	1	2	ct	14	37	64	Non-coding
CcC01229_1	1	3	cta	9	189	215	Non-coding
CcC01229_1	2	3	ggt	6	265	282	Non-coding
CcC01247_1	1	3	cca	7	337	357	Coding
CcC01249_1	1	2	tc	50	9	108	Non-coding
CcC01249_2	1	2	tc	54	232	339	Unknown
CcC01257_1	1	4	tctt	6	220	243	Non-coding
CcC01271_1	1	2	tc	11	9	30	Non-coding
CcC01271_1	2	3	cag	6	497	514	Coding
CcC01279_1	1	3	tca	8	240	263	Non-coding
CcC01298_1	1	3	tct	8	52	75	Non-coding
CcC01303_1	1	3	agc	6	300	317	Coding
CcC01314_1	1	2	tc	11	9	30	Non-coding
CcC01316_1	1	2	ct	9	46	63	Non-coding

Supplementary Table IV. Continued.

Potential unigene ID	SSR number	Motif length	Motif	Number of motif	Stat position	End position	SSR location
CcC01317_1	1	2	tc	10	50	69	Non-coding
CcC01317_2	1	2	tc	10	39	58	Non-coding
CcC01317_2	2	3	gaa	8	252	275	Coding
CcC01321_1	1	3	gga	6	300	317	Coding
CcC01335_1	1	2	ga	14	186	213	Non-coding
CcC01358_1	1	2	tc	9	167	184	Unknown
CcC01361_1	1	3	ttc	9	347	373	Unknown
CcC01400_1	1	3	ggc	6	449	466	Coding
CcC01401_1	1	3	gga	6	324	341	Coding
CcC01410_1	1	3	caa	6	323	340	Coding
CcC01420_1	1	2	ag	13	176	201	Non-coding
CcC01426_1	1	2	ct	9	31	48	Non-coding
CcC01428_1	1	2	ct	57	9	122	Unknown
CcC01435_1	1	2	tc	10	78	97	Non-coding
CcC01442_1	1	2	ct	17	14	47	Non-coding
CcC01442_1	2	2	tc	9	74	91	Non-coding
CcC01451_1	1	3	cca	9	171	197	Coding
CcC01462_1	1	2	ct	18	9	44	Non-coding
CcC01471_1	1	2	tc	14	69	96	Non-coding
CcC01479_1	1	2	ca	9	50	67	Non-coding
CcC01500_1	1	2	ct	18	37	72	Unknown
CcC01500_1	2	2	tc	13	249	274	Unknown
CcC01504_1	1	2	ct	9	312	329	Non-coding
CcC01513_1	1	3	cag	6	251	268	Coding
CcC01546_1	1	2	tc	13	131	156	Non-coding
CcC01548_1	1	2	ag	16	71	102	Unknown
CcC01564_1	1	3	ctc	8	50	73	Non-coding
CcC01568_1	1	3	cca	6	222	239	Coding
CcC01593_1	1	3	caa	8	91	114	Coding
CcC01604_1	1	2	ag	22	278	321	Unknown
CcC01611_1	1	3	gca	8	39	62	Coding
CcC01614_1	1	3	cca	6	316	333	Coding
CcC01619_1	1	2	tc	11	24	45	Non-coding
CcC01640_1	1	2	ct	10	500	519	Unknown
CcC01650_1	1	2	at	14	276	303	Unknown
CcC01653_1	1	3	acc	7	388	408	Unknown
CcC01662_1	1	3	ata	9	177	203	Non-coding
CcC01671_2	1	3	atg	7	98	118	Coding
CcC01672_1	1	2	ct	12	27	50	Non-coding
CcC01672_1	2	2	ac	12	183	206	Non-coding
CcC01681_1	1	2	ga	9	10	27	Non-coding
CcC01686_1	1	2	ct	13	74	99	Non-coding
CcC01687_1	1	3	cca	7	257	277	Coding
CcC01707_1	1	2	ct	13	49	74	Unknown
CcC01727_1	1	3	act	6	207	224	Coding
CcC01736_1	1	3	gga	6	324	341	Coding
CcC01739_1	1	3	gca	7	292	312	Coding
CcC01748_1	1	2	tc	17	153	186	Unknown
CcC01767_1	1	2	ct	11	24	45	Non-coding
CcC01772_1	1	2	ag	12	105	128	Non-coding
CcC01777_1	1	3	cgc	7	475	495	Coding
CcC01805_1	1	3	atc	6	37	54	Coding
CcC01814_1	1	3	ccg	8	74	97	Coding
CcC01821_1	1	2	ca	17	17	50	Unknown
CcC01824_1	1	2	ct	11	9	30	Non-coding
CcC01851_1	1	3	acc	8	163	186	Coding
CcC01851_1	2	3	aac	6	319	336	Coding

Supplementary Table IV. Continued.

Potential unigene ID	SSR number	Motif length	Motif	Number of motif	Stat position	End position	SSR location
CcC01857_1	1	3	tct	9	201	227	Non-coding
CcC01865_1	1	2	tc	10	46	65	Non-coding
CcC01873_1	1	2	ct	16	10	41	Non-coding
CcC01875_1	1	3	gag	7	145	165	Coding
CcC01879_1	1	2	ct	13	11	36	Non-coding
CcC01882_1	1	3	ttc	6	173	190	Non-coding
CcC01890_1	1	3	tgc	6	191	208	Coding
CcC01899_1	1	3	gaa	6	61	78	Coding
CcC01909_1	1	3	gga	6	315	332	Coding
CcC01916_1	1	2	tc	12	8	31	Non-coding
CcC01931_1	1	2	ag	9	34	51	Non-coding
CcC01934_1	1	3	gcc	6	240	257	Coding
CcC01952_1	1	2	tc	13	101	126	Non-coding
CcC01954_1	1	2	tc	18	66	101	Non-coding
CcC01967_1	1	2	ct	11	88	109	Non-coding
CcC01967_1	2	3	gtt	6	50	67	Non-coding
CcC01969_1	1	3	cca	6	302	319	Coding
CcC01989_1	1	3	tgt	7	110	130	Non-coding
CcC01991_1	1	2	at	9	628	645	Non-coding
CcC01997_1	1	2	ct	13	75	100	Unknown
CcC01999_1	1	2	tc	16	141	172	Non-coding
CcC02014_1	1	2	tc	11	37	58	Non-coding
CcC02014_1	2	3	aag	7	77	97	Non-coding
CcC02022_1	1	2	ct	12	159	182	Non-coding
CcC02023_1	1	2	ct	11	27	48	Non-coding
CcC02027_1	1	2	tc	17	48	81	Non-coding
CcC02031_1	1	3	gtc	7	550	570	Coding
CcC02044_1	1	2	gt	9	93	110	Non-coding
CcC02044_1	2	2	ga	16	111	142	Non-coding
CcC02053_1	1	2	ag	10	49	68	Non-coding
CcC02057_1	1	2	ta	18	637	672	Non-coding
CcC02065_1	1	2	ct	11	91	112	Non-coding
CcC02069_1	1	3	caa	6	177	194	Coding
CcC02075_1	1	2	ct	10	22	41	Non-coding
CcC02084_1	1	2	ct	9	7	24	Non-coding
CcC02097_1	1	2	tc	15	34	63	Non-coding
CcC02103_1	1	2	ta	16	189	220	Non-coding
CcC02107_1	1	3	ggt	10	481	510	Coding
CcC02111_1	1	2	tg	9	36	53	Non-coding
CcC02112_1	1	2	ta	16	52	83	Unknown
CcC02123_1	1	2	ct	10	30	49	Non-coding
CcC02162_1	1	2	ct	10	16	35	Non-coding
CcC02171_1	1	2	ct	10	1	20	Unknown
CcC02182_1	1	3	gga	6	314	331	Coding
CcC02186_1	1	3	tct	9	73	99	Non-coding
CcC02187_1	1	2	ag	20	163	202	Unknown
CcC02197_1	1	2	tc	14	281	308	Unknown
CcC02197_1	2	3	ttc	7	13	33	Unknown
CcC02227_1	1	2	tc	9	40	57	Non-coding

Supplementary Table IV. Continued.

Potential unigene ID	SSR number	Motif length	Motif	Number of motif	Stat position	End position	SSR location
CcC02238_1	1	2	at	12	66	89	Unknown
CcC02268_1	1	2	ta	14	533	560	Non-coding
CcC02291_1	1	3	cag	9	262	288	Coding
CcC02294_1	1	3	aag	6	75	92	Non-coding
CcC02296_1	1	2	ct	15	53	82	Non-coding
CcC02312_1	1	3	ttc	6	329	346	Non-coding
CcC02315_1	1	2	tc	18	10	45	Non-coding
CcC02329_1	1	3	gat	7	220	240	Coding
CcC02337_1	1	2	ct	13	99	124	Non-coding
CcC02358_1	1	2	ag	11	78	99	Unknown
CcC02375_1	1	3	gaa	7	180	200	Coding
CcC02379_1	1	2	ct	13	17	42	Non-coding
CcC02384_1	1	2	ct	10	11	30	Unknown
CcC02386_1	1	3	gat	7	320	340	Coding
CcC02391_1	1	3	gga	6	291	308	Coding
CcC02406_1	1	3	tta	8	158	181	Non-coding
CcC02408_1	1	2	ct	11	358	379	Unknown
CcC02416_1	1	2	tc	14	18	45	Non-coding
CcC02418_1	1	2	at	10	639	658	Non-coding
CcC02419_1	1	2	ga	10	480	499	Non-coding
CcC02420_1	1	2	ga	13	71	96	Non-coding
CcC02438_1	1	2	ct	10	34	53	Non-coding
CcC02444_1	1	2	tc	11	48	69	Non-coding
CcC02448_1	1	3	aag	9	322	348	Unknown
CcC02450_1	1	2	ct	10	9	28	Non-coding
CcC02464_1	1	2	tc	20	71	110	Non-coding
CcC02489_1	1	2	tc	14	62	89	Non-coding
CcC02489_1	2	2	tc	19	138	175	Non-coding
CcC02496_1	1	2	tc	10	34	53	Non-coding
CcC02500_1	1	3	taa	6	290	307	Non-coding
CcC02501_1	1	3	aca	7	179	199	Coding
CcC02510_1	1	2	ag	9	267	284	Unknown
CcC02512_1	1	3	gga	7	257	277	Coding
CcC02524_1	1	3	gcc	6	302	319	Coding
CcC02535_1	1	2	at	12	306	329	Non-coding