

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

**In silico mining, characterization and cross-species transferability of EST-SSR markers for European hazelnut (*Corylus avellana* L.)**

**This is the author's manuscript**

*Original Citation:*

*Availability:*

This version is available <http://hdl.handle.net/2318/1622333> since 2017-01-18T15:40:33Z

*Published version:*

DOI:10.1007/s11032-015-0195-7

*Terms of use:*

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

1 ***In silico* mining, characterization and cross-species transferability of EST-SSR**  
2 **markers for European hazelnut (*Corylus avellana* L.)**

3  
4 Boccacci P.<sup>1,2,\*</sup>, Sandoval Prando M.A.<sup>2,\*</sup>, Beltramo C.<sup>2)</sup>, Lembo A.<sup>3)</sup>, Sartor C.<sup>2)</sup>, Mehlenbacher S.A.<sup>4)</sup>,  
5 Botta R.<sup>2)</sup>, Torello Marinoni D.<sup>2)</sup>

6  
7 \* These authors contributed equally to this work

8  
9 <sup>1)</sup> Institute for Sustainable Plant Protection - National Research Council (IPSP-CNR), Unit of Grugliasco, Largo Paolo  
10 Braccini, 2 - 10095 Grugliasco (TO), Italy

11 <sup>2)</sup> Department of Agriculturae, Forestry and Food Science (DISAFA) - University of Turin, Largo Paolo Braccini, 2 -  
12 10095 Grugliasco (TO), Italy

13 <sup>3)</sup> Molecular Biotechnology Center, Department of Molecular Biotechnology and Health Sciences, University of Turin,  
14 Via Nizza 52 – 10126 Torino, Italy

15 <sup>4)</sup> Department of Horticulture, Oregon State University, Agricultural and Life Sciences Bldg. 4017, Corvallis, OR  
16 97331, USA

17

18

19

20

21

22

23

24

25 Corresponding author: Paolo Boccacci, e-mail: p.boccacci@ivv.cnr.it, phone: +39.011.411.7304, fax:  
26 +39.011.6708658

27

28 **Abstract**

29 The European hazelnut (*Corylus avellana* L.) is one of the most important nut crops. In this work we  
30 characterize functional microsatellite or simple sequence repeat (SSR) markers for genetic analysis and  
31 molecular breeding in this species. A total of 38,454 Betulaceae EST sequences from NCBI resulted in  
32 1,282 non-redundant EST-SSRs. Dinucleotide repeats were the most abundant (63.9%), followed by  
33 trinucleotides (33.8%). The putative functions of the non-redundant EST-SSRs were classified according to  
34 gene ontology (GO) categories (biological process, molecular function, and cellular component). A total of  
35 921 sequences showed significant hits with the non-redundant protein database, and GO categories were  
36 assigned to 696 (75.5%) of them. Flanking primer pairs were designed for 78 di- and trinucleotide EST-  
37 SSRs from *Alnus glutinosa* L. (29), *Betula pendula* Roth (26), and *Betula platyphylla* Suckaczev (23).  
38 Further, 41 dinucleotide repeats selected from hazelnut transcriptome sequences were added. Thirty-six out  
39 119 primer pairs generated amplification products in six hazelnut accessions and in the samples of the  
40 species from which they were isolated. Among them, 20 were polymorphic when tested on 18 hazelnut  
41 cultivars. Fifteen loci are suitable for mapping in a F<sub>1</sub> population of 'Tonda Gentile delle Langhe' x  
42 'Merveille de Bollwiller' and 11 of them were functionally annotated. The cross-species transferability of  
43 36 EST-SSR loci within nine *Corylus* species was also performed. The success rate of markers  
44 transferability (including *C. avellana*) ranged from 11% to 100%, with an average of 55%. The EST-SSRs  
45 developed increase the number of markers currently available for hazelnut.

46  
47  
48 **Key words:** expressed sequence tags; microsatellite; functional annotation; polymorphism; transferability;  
49 filbert

50  
51

## 52 **Introduction**

53 Betulaceae, one of eight families of the Order Fagales, includes six living genera and about 140 species. It  
54 is subdivided into two clades: Betuloideae, which includes the genera *Alnus* (35 species) and *Betula* (35–60  
55 species), and Coryloideae, which contains *Carpinus* (35 species), *Corylus* (11–13 species), *Ostrya* (10  
56 species), and *Ostryopsis* (2 species) (Chen et al. 1999; Yoo and Wen 2002). Except for *Ostryopsis*, which  
57 is endemic to eastern Asia, the other five genera show similar patterns of distribution throughout the  
58 Northern Hemisphere. The basic chromosome number is  $n = 14$  for *Alnus* and *Betula*, 11 for *Corylus*, and  
59 8 for *Carpinus*, *Ostrya*, and *Ostryopsis*. Several species of *Betula* form a polyploid series, with chromosome  
60 numbers of  $2n = 28, 56, 70, 84,$  and 112 (Järvinen et al. 2004).

61 The *Corylus* genus comprises from 9 to 25 species, depending on the authority. Recent revisions  
62 based on morphological, molecular, and hybridization studies suggest around eleven polymorphic species  
63 assigned to two sections: *Acanthochlamys* and *Corylus* (Erdoğan and Mehlenbacher 2000; Whitcher and  
64 Wen 2001). Section *Acanthochlamys* includes only *C. ferox* Wall, which has a spiny chestnut-like involucre  
65 unlike any other species in the genus. Section *Corylus* traditionally includes three subsections:  
66 *Phyllochlamys* contains three shrubs species with leafy involucre (*C. avellana* L., *C. americana* Marshall,  
67 and *C. heterophylla* Fisch.); *Siphonochlamys* includes three bristle-husked shrubs species (*C. cornuta*  
68 Marshall, *C. californica* Marshall, and *C. sieboldiana* Blume.); *Columnaea* includes four tree species (*C.*  
69 *columna* L., *C. chinensis* Franch., *C. jacquemontii* Decne., and *C. fargesii* C.K. Schneider). The most widely  
70 known and well-studied member, the European hazelnut (*C. avellana* L.), is one of the most important nut  
71 crops in terms of worldwide production. The Black Sea countries account for the majority of world  
72 production (2008–2012): Turkey (598,158 tons), Azerbaijan (30,030 tons), and Georgia (25,020 tons). Other  
73 important producers are Italy (104,577 tons), the USA (32,399 tons), Iran (20,832 tons), China (19,700  
74 tons), and Spain (16,239 tons), with significant new plantings in Chile (FAOstat 2014). About 90% of the  
75 world crop is shelled and sold as kernels, while the remaining 10% is sold in-shell for fresh consumption.  
76 The primary user of kernels, the food industry, requires cultivars that produce nuts with few defects and has  
77 precise requirements for morphological, chemical, and physical characteristics of the kernels. *C. avellana*

78 also possesses many attributes that make it an attractive candidate for use as a model system for the  
79 Betulaceae family. This species has a relatively short life cycle, bearing seeds at around 5 year, has a short  
80 stature for a tree (~5 m), a small genome that, at ~400 Mb, is about triple that of the established dicot plant  
81 model *Arabidopsis thaliana* (L.) Heynh. (125 Mb).

82         Microsatellites or simple sequence repeats (SSRs) are tandemly repeated 1–6 bp sequence motifs.  
83 They have many characteristics of the ideal molecular marker: multi-allelic nature, co-dominant inheritance,  
84 reproducibility, high polymorphism, transferability to related species and genera, relative abundance and  
85 good genome coverage (Powell et al. 1996). SSRs are divided in two categories: genomic SSRs (gSSRs),  
86 derived from random genomic sequences, and genic SSRs or EST-SSRs, derived from expressed sequence  
87 tags. Generally, gSSRs have neither genic function nor close linkage to transcriptional regions, while EST-  
88 SSRs are potentially located within genes of known or at least putative function that may control some  
89 agronomic traits. EST-SSRs tend to be more readily transferable between related species or genera than  
90 genomic ones, since coding sequences are better conserved than non-coding sequences. However, because  
91 of lower polymorphism, they are not as efficient as gSSRs for distinguishing closely related genotypes. On  
92 the other hand, EST-SSRs are powerful tools for linkage map construction, comparative mapping, marker-  
93 assisted selection, and evolutionary studies (Varshney et al. 2005). The use of microsatellites was limited in  
94 plants by the costs involved in isolating large numbers of SSRs from the target species. Although the  
95 isolation of microsatellites by the enrichment procedure (Edwards et al. 1996) is now more efficient than in  
96 the past, *in silico* mining of SSRs from sequence databases provides a time- and cost-effective alternative.  
97 With the development of next-generation sequencing techniques, more and more EST-SSRs have been  
98 deposited in several databases. Among the tree plant species, EST-SSR markers were mined and  
99 characterized in apple (Gasic et al. 2009), apricot (Decroocq et al. 2003), *Citrus* spp. (Chen et al. 2006),  
100 eucalypts (Acuña et al. 2012), kiwifruit (Fraser et al. 2004), grape (Decroocq et al. 2003; Huang et al. 2011),  
101 and rubber tree (Feng et al. 2009) among others.

102         In *C. avellana*, more than 230 gSSR loci have been developed (Bassil et al. 2005a, 2005b, 2013;  
103 Boccacci et al. 2005; Gürcan and Mehlenbacher 2010a, 2010b; Gürcan et al. 2010a). SSRs have been

104 extensively used to characterize genetic variation in hazelnut germplasm (Boccacci and Botta 2010;  
105 Boccacci et al. 2006, 2008, 2013; Campa et al. 2011; Gökirmak et al. 2009; Gürcan et al. 2010b) and in  
106 *Corylus* spp. (Sathuvalli and Mehlenbacher 2012; Bassil et al. 2013). Moreover, SSRs have been placed on  
107 linkage maps (Mehlenbacher et al. 2006; Beltramo et al. 2012) and different sources of eastern filbert blight  
108 (EFB) resistance have been assigned to linkage groups based on co-segregation with mapped SSRs  
109 (Sathuvalli et al. 2011). On the contrary, only 10 EST-SSR from *Betula pendula* Roth have been  
110 characterized in *Corylus* spp. (Gürcan and Mehlenbacher 2010a), while an initial whole-transcriptome  
111 assembly was recently completed in *C. avellana* from the cultivar 'Jefferson' (Rowley et al. 2012).

112         The present study aims to develop conserved orthologous markers for genetic analysis of different  
113 Betulaceae species and to expand the genomic resources for the European hazelnut breeding and mapping.  
114 To achieve this, an *in silico* identification and characterization of unique SSRs derived from Betulaceae  
115 ESTs, retrieved from a public database, was performed. Microsatellites were identified in cDNA sequences  
116 of black alder (*Alnus glutinosa* L.), silver birch (*Betula pendula* Roth), and Japanese white birch (*Betula*  
117 *platyphylla* Suckaczev). In addition, SSRs were selected from *C. avellana* transcriptome sequences. A set  
118 of EST-SSR markers was developed and evaluated for their ability to detect polymorphism in hazelnut,  
119 suitability for genetic mapping in a F<sub>1</sub> full-sib progeny of 'Tonda Gentile delle Langhe' x 'Merveille de  
120 Bollwiller', and transferability across the genus *Corylus*.

121

## 122 **Materials and methods**

### 123 *Mining of SSR-containing sequences and functional annotation*

124 All Betulaceae EST sequences were downloaded from the EST database (dbEST) of NCBI GenBank  
125 (<http://www.ncbi.nlm.nih.gov/dbEST/>) in January 2011. Among a total of 38,454 ESTs, 32,544 were from  
126 *A. glutinosa*, 2,549 from *B. pendula*, and 3,361 from *B. platyphylla*.

127         SSR-containing sequences were identified using MISA software (Thiel et al. 2003), a Perl script  
128 which allows both perfect and compound SSRs to be detected. As mononucleotide repeat polymorphisms  
129 are often difficult to interpret, only di-, tri-, tetra-, penta-, and hexanucleotide motifs were considered as

130 potential candidates for EST-SSR marker development. The minimal length of SSR repeats was defined as  
131  $2 \times 7 = 14$  bp for dinucleotides,  $3 \times 5 = 15$  bp for trinucleotides,  $4 \times 5 = 20$  bp for tetranucleotides,  $5 \times 5 =$   
132  $25$  bp for pentanucleotides, and  $6 \times 5 = 30$  bp for hexanucleotides. For compound repeats, the maximum  
133 default interruption (spacer) length was set at 100 bp. A ClustalW v. 2.1 multiple sequence alignment  
134 (Larkin et al. 2007) was performed among the SSR-containing ESTs to identify non-redundant sequences.

135 Functional annotation of EST-SSR sequences was performed using a Blast2GO workstation (Conesa  
136 et al. 2005). BLASTx searches were performed against the NCBI non-redundant protein database  
137 (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) with an e-value cutoff  $\leq 1e^{-5}$ . Gene Ontology (GO) terms were  
138 assigned to the hits obtained after the BLASTx search according to the GO consortium (Ashburner et al.  
139 2000). Enzyme codes (EC) were acquired by mapping from equivalent GOs. The Kyoto Encyclopedia of  
140 Genes and Genomes (KEGG) map module was used to display the enzymatic functions in the context of the  
141 metabolic pathways in which they participate.

142  
143 *Plant materials and DNA extraction*  
144 PCR amplification was initially tested on 6 *C. avellana* accessions ('Tonda Gentile delle Langhe',  
145 'Merveille de Bollwiller', 'Culplà', 'Negret', 'Imperial de Trebizonde', and 'Tombul') and 2 individuals  
146 each of *A. glutinosa*, *B. pendula* and *B. platyphylla*. SSR polymorphism and transferability were assessed  
147 on a total of 34 hazelnut accessions, including 18 *C. avellana* cultivars and 15 accessions representing 9  
148 *Corylus* species (Online data 1). *C. avellana* cultivars were chosen to represent the various countries that  
149 grow hazelnut, including Turkey, Italy, Spain, and the United States. 'Tonda Gentile delle Langhe' (TGL)  
150 and 'Merveille de Bollwiller' (syn. 'Hall's Giant') are also the parents of a full-sibling progeny obtained  
151 from a controlled cross and used to construct a genetic linkage map by Beltramo et al. (2012). DNA was  
152 extracted from 0.2 g of leaves using a modified procedure described by Thomas et al. (1993).

153 *PCR amplification and fragment analysis*  
154 Sequences flanking the microsatellite motifs in 119 ESTs were used to design primer pairs using Primer3  
155 software (Rozen and Skaletsky 2000). Seventy-nine were from the Betulaceae dbEST and 41 from the *C.*

156 *avellana* transcriptome sequences. The hazelnut transcriptome consists of 28,255 contigs  
157 (<http://hazelnut.cgrb.oregonstate.edu>) that were sequenced, assembled, and functionally annotated by  
158 Rowley et al. (2012) from a diverse set of tissues and organ type (young leaves, catkins, bark, and whole  
159 young seedlings). Primers were designed with the following criteria: i) melting temperature ( $T_m$ ) between  
160 57°C and 63°C; ii) product size of 100 bp to 300 bp; iii) primer length of 18 bp to 20 bp with amplification  
161 rate  $\geq$  80%; iv) GC% content between 40% and 80%. A 20-bp long universal M13 forward primer sequence  
162 5'-CAC GAC GTT GTA AAA CGAC- 3' (Zhang et al. 2003) was added as a common tail to the 5' end of  
163 all 119 SSR forward primers. The universal M13 primers were labeled with a fluorochrome (6-FAM, HEX,  
164 NED, or PET).

165 PCR amplifications were performed in a volume of 15  $\mu$ l containing 40 ng DNA, 1x Biolase  $\text{NH}_4$   
166 reaction buffer [160 mM  $(\text{NH}_4)_2\text{SO}_4$ , 670 mM Tris-HCl (pH 8.8 at 25°C), 0.1% Tween-20], 2.25 mM  
167  $\text{MgCl}_2$ , 200  $\mu$ M of each dNTPs, 0.75  $\mu$ l of a primer mix containing 8 pmol of reverse primer and 2 pmol of  
168 the forward primer, 1.2  $\mu$ l of 5 pmol M13 tail primer, and 0.4 U Biolase DNA polymerase 5U/ $\mu$ l (Bioline,  
169 London, UK). The PCR conditions were: a first denaturation step at 94°C for 4 min, then 30 cycles of  
170 denaturation (45 s at 94 °C), annealing (45 s at the optimal temperature for each primer pair, as reported in  
171 Table 1), and extension (45 s at 72°C), followed by 10 cycles of 94°C for 45 s, 53-54°C (Table 1) for 45 s,  
172 and 72°C for 45 s. The final elongation step was at 72°C for 10 min.

173 Amplification products were initially separated by electrophoresis on 3% agarose gels by staining  
174 with GelRed™ Nucleic Acid Gel Stain (Biotium, Hayward, CA, USA). Loci that showed good amplification  
175 were then analyzed using an ABI-PRISM 3130 Genetic Analyzer capillary electrophoresis instrument  
176 (Applied Biosystems, Foster City, CA, USA). Results were processed with GeneMapper software and  
177 alleles were designated by their size in base pairs (bp) using a GeneScan-500 LIZ standard (Applied  
178 Biosystems).

#### 179 *Data analysis*

180 Microsatellite data obtained for 18 *C. avellana* cultivars were processed using the PowerMarker software  
181 (Liu and Muse 2005). The number of alleles ( $N_A$ ), number of genotypes ( $N_G$ ), genetic diversity ( $H_e$ ),



182 observed heterozygosity ( $H_o$ ), and polymorphic information content (PIC) were calculated at each locus.  $H_e$   
183 was computed as  $H_e = 1 - \sum p_i^2$ , where  $p_i$  is the frequency of the  $i^{\text{th}}$  allele.  $H_o$  was from direct counts. PIC  
184 values of each locus was estimated as  $PIC = 1 - \sum p_i^2 - \sum p_i^2 p_j^2$ , where  $p_i$  and  $p_j$  are the frequencies of the  $i^{\text{th}}$   
185 and  $j^{\text{th}}$  alleles, respectively (Botstein et al. 1980).

186

## 187 **Results and discussion**

### 188 *Identification, frequency and characterization of EST-SSRs*

189 The SSR screening with MISA software of 38,454 Betulaceae EST sequences identified 3,950 that  
190 contained SSRs. Nevertheless, the SSR number may be overestimated because of the redundancy of ESTs  
191 from Genbank. In order get non-redundant data, ClustalW v. 2.1 software was used to cluster the 3,950  
192 EST-SSRs and a total of 1,282 non-redundant sequences were identified. These results revealed a 3.1-fold  
193 redundancy among the EST-SSRs from Betulaceae, which was similar to what was found by Huang et al  
194 (2011) in *Vitis* spp. (3.8-fold). A total of 1,420 SSRs from the 1,282 non-redundant ESTs were identified  
195 by MISA; 123 ESTs (9.6%) contained more than one SSR. Among the 1,420 SSRs, 1,320 (93%) had simple  
196 repeat motifs while 100 (7%) were compound types.

197 A total of 38 different SSR motif types were present (Online data 2). Dinucleotide (63.9%) repeats  
198 were the most abundant and the AG/CT motif was the most common (79.9%), followed by AT/AT (16.1%),  
199 AC/GT (3.9%), and CG/CG (0.1%). Trinucleotides accounted for 33.8% and all ten possible tri- motifs were  
200 recovered, of which AAG/CTT was the most frequent (39%). Tetra-, penta- and hexanucleotide repeats  
201 together represented 2.3% of the total non-redundant EST-SSRs. Di- and trinucleotide SSR loci were  
202 predominant (97.7%) in the Betulaceae dbEST. In other plants species, either di- (e.g., Feng et al. 2009) or  
203 trinucleotide SSRs (e.g., Acuña et al. 2012; Chen et al. 2006; Huang et al. 2011) have been reported as most  
204 common. Trinucleotide repeats are predominant compared with other motifs in coding sequences. This high  
205 frequency is likely due to a selective disadvantage of non-trimeric SSR variants in coding regions and the  
206 resulting frameshift mutations (Metzgar et al. 2000). Dinucleotide repeats are typically more frequent in 5'-  
207 and/or 3'-UTR regions, but occasionally occur in coding regions (Varshney et al. 2005). Among them, high

208 proportions of AG/CT repeats in ESTs were found in rubber tree (83.7%) (Feng et al. 2009) and kiwifruit  
209 (70%) (Fraser et al. 2004). An abundance of AG/CT repeats seems to be a general property of plant genomes,  
210 whereas AC/GT repeats are more common in animal genomes. The functional significance of SSRs in genic  
211 regions of plants is unclear, but homopurine–homopyrimidine stretches like AG/CT in the 5'-UTR have  
212 been reported to be involved in gene regulation (Varshney et al. 2005).

213

#### 214 *Functional annotation of ESTs containing SSRs*

215 The 1,282 non-redundant EST-SSR sequences from Betulaceae were used in an initial BLAST search  
216 against the NCBI non-redundant protein database. A total of 921 (71.8%) sequences showed 17,390  
217 significant hits. Among them, 9.7% of the total hits matched homologous soybean (*Glycine max* L.)  
218 sequences, 8.9% matched both grape (*Vitis vinifera* L.) and poplar (*Populus trichocarpa* Torr. and A. Gray)  
219 sequences, 7% matched cacao (*Theobroma cacao* L.) sequences, and 56.8% matched sequences from other  
220 crop plant species. No BLAST hits were observed for 361 sequences. Similar results were reported for 4,909  
221 EST-SSRs from moss [*Physcomitrella patens* (Hedw.) Bruch and Schimp.] by Victoria et al. (2011). Of the  
222 total hits, they found matches of 10.8% and 8.6% to grape and poplar, respectively, while only 3.9%  
223 corresponded to soybean. Also in the functional annotation of the hazelnut transcriptome by Rowley et al.  
224 (2012), a high similarity of the proteins encoded by the transcript contigs was observed with grape (36.6%)  
225 and poplar (23.1%).

226 Functional annotations were performed on 921 EST-SSRs using the program Blast2GO, a tool for  
227 assigning GO terms to unknown sequences (Conesa et al. 2005). This resulted in GO functional  
228 classifications for 696 (75.5%) sequences comprising a total of 3,663 GO annotations, which allowed  
229 assignment to one of three functional categories: biological process, molecular function, and cellular  
230 component (Fig. 1). Regarding biological process (Fig 1A), 560 consensus sequences were subdivided in  
231 11 subcategories, among which the cellular (79.3%) and metabolic (78%) processes were the most common.  
232 Considering molecular functions (Fig 1B), 509 ESTs were assigned to two subcategories: binding (64.4%)  
233 and catalytic activity (82.9%). When grouped according to the cellular component category (Fig 1C), 494

234 sequences were assigned to one of four subcategories, with 93.3% and 72% of them in the subcategories  
235 cell and organelle, respectively. GO terms at higher and lower levels for each of the three main GO  
236 categories are reported in Online data 3. EC number and KEGG maps are described in Online data 4.  
237 Relevant cellular metabolic processes were fully represented by at least one EC number and the EST-SSRs  
238 annotated involved a total of 93 metabolic pathways, including those involved in the biosynthesis of  
239 unsaturated or saturated fatty acids. In hazelnut, kernels have a high content of fat ( $\geq 60\%$ ), and this is one  
240 of the determinants of kernel flavor. Both lipid content and the proportion of fatty acids are important criteria  
241 in the evaluation of kernel quality. Moreover, a high concentration of unsaturated fatty acids and the  
242 presence of natural antioxidants ( $\alpha$ -tocopherol) and pro-oxidant minerals (iron and copper) are factors  
243 involved in kernel rancidity. Therefore, cultivars rich in anti-oxidants but with a low unsaturated/saturated  
244 ratio, low pro-oxidant compound content, and low enzymatic activity, are preferred. In fact this combination  
245 of traits would minimize post-harvest quality losses, and packaging and refrigeration costs (Bacchetta et al.  
246 2013).

247

#### 248 *EST-SSR validation and polymorphism in C. avellana*

249 A set of 78 EST-SSRs were selected among the non-redundant di- and trinucleotide microsatellites retrieved  
250 from the Betulaceae ESTs. Among them, 29 were from *A. glutinosa*, 26 from *B. plathyphylla*, and 23 from  
251 *B. pendula*. An additional 41 dinucleotide EST-SSRs were selected from the *C. avellana* transcriptome. All  
252 sequences contained a single perfect microsatellite repeat and sufficient flanking regions of appropriate  
253 quality for primer design.

254 The amplification ability of the 119 primer pairs was initially performed using six *C. avellana*  
255 cultivars and two accessions each of *A. glutinosa*, *B. plathyphylla*, and *B. pendula*. As evidenced by  
256 electrophoresis on 3% agarose gels, 21 of the 29 (72.4%) *Alnus* primer pairs, 20 of the 26 (76.9%) *B.*  
257 *plathyphylla* primers pairs, 13 of the 23 (56.5%) *B. pendula* primers pairs, and 12 of the 41 (29.3%) *Corylus*  
258 primer pairs generated PCR products in the expected size range in the samples of the respective species.  
259 These results indicated a higher amplification level in *Alnus* and *Betula*, compared to that obtained in

260 *Corylus*. However, the main purpose of this work was to identify EST-SSR markers from Betulaceae for  
261 European hazelnut. For this reason, we considered only loci that generated one or two distinct PCR products  
262 in the expected size range both in all hazelnut accessions and in their control samples (i.e. species of EST  
263 origin). This goal was reached by 36 out of 119 (30.2%) primer pairs, 12 from *A. glutinosa*, 7 from *B.*  
264 *plathyphylla*, 5 from *B. pendula*, and 12 from *C. avellana* (Table 1 and Online data 5). As expected,  
265 amplifications were better in species of the genus from which the EST-SSRs were developed than in the  
266 other genera. Gasic et al. (2009) reported that the highest amplification of apple (*Malus x domestica* Borkh)  
267 EST-SSRs across individual Rosaceae species was 62% in the closely related pear (*Pyrus communis* L.),  
268 whereas 38% and 37% amplified alleles in peach [*Prunus persica* (L.) Batsch] and almond [*Prunus dulcis*  
269 (Mill.) D.A. Webb], respectively, and 28% amplified in the genus *Rosa*. Heesacker et al. (2008) also  
270 reported that 88.6% of the 466 sunflower (*Helianthus annuus* L.) EST-SSR or InDel markers amplified  
271 alleles from one or more wild species, while 14.8% amplified in related safflower (*Carthamus tinctorius*  
272 L.), and 14.4% amplified in lettuce (*Lactuca sativa* L.), a distantly related genus in the Asteraceae family.  
273 About 70% of the primer pairs gave no amplification product in hazelnut. Amplification failures could be  
274 due to several factors, such as primers extending across splicing sites, presence of large introns in the  
275 genomic sequences, low quality of the EST sequences, or primer sequences derived from chimeric cDNA  
276 clones.

277 Thirty-six loci were fluorescently labelled and selected in order to evaluate their polymorphism in 18  
278 *C. avellana* cultivars; PCR products were analyzed using a capillary electrophoresis instrument. Primer  
279 pairs amplified in all 18 hazelnut samples (Online data 6) and 20 loci (55.6%) were polymorphic, while 16  
280 were monomorphic. Characterization data for the 20 polymorphic loci are summarized in Table 2. A total  
281 of 92 alleles was observed and the  $N_A$  per locus ranged from 2 (Corav692, Corav2241, and Corav2564) to  
282 8 (Ag4395 and Corav2560), with a mean of 4.6. The  $N_G$  ranged from 2 (Corav692 and Corav2241) to 14  
283 (Corav2560), with a mean value of 5.90.  $H_e$  averaged 0.56 and ranged from 0.15 (Corav2241) to 0.83  
284 (Corav2560), while  $H_o$  averaged 0.57 and ranged from 0.17 (Corav2241) to 1.00 (Corav4911). PIC ranged  
285 from 0.14 for Corav2241 to 0.81 for Corav2560 and its mean value was 0.50. PIC values were  $< 0.50$  for

286 11 loci and ranged from 0.51 to 0.81 for the 9 remaining loci, which were highly informative. The level of  
287 polymorphism obtained was similar to or slightly higher than reported in comparable studies. At the 18  
288 EST-SSR loci investigated by Wöhrmann and Weising (2011) in 12 pineapple [*Ananas comosus* (L.) Merr.]  
289 cultivars, 77 alleles were observed. The number of alleles ranged from 2 to 6 per locus (mean 4.3), average  
290  $H_e$  and  $H_o$  were 0.58 and 0.49, respectively. Feng et al. (2009) developed 30 primer pairs in rubber tree  
291 (*Hevea brasiliensis*), and amplification of 12 cultivars produced an average of 2.47 alleles per locus and an  
292 average PIC of 0.38. Varshney et al. (2005) reported that the average PIC ranged to 0.32 to 0.66 in different  
293 herbaceous species, such as barley, coffee, rice, sugarcane, and wheat. On the contrary, our EST-SSR  
294 revealed less polymorphism in comparison to genomic SSRs used in hazelnut germplasm characterization.  
295 Bassil et al. (2005a) analyzed 25 gSSR loci in 20 genotypes and mean  $N_A$ ,  $H_o$ ,  $H_e$ , and PIC were 7.16, 0.62,  
296 0.68, and 0.64, respectively. Boccacci et al. (2005) developed 18 gSSR and showed a high level of  
297 polymorphism in 20 accessions, with mean  $N_A$ ,  $H_e$ ,  $H_o$ , and PIC of 7.10, 0.67, 0.70, and 0.64, respectively.  
298 The lower polymorphism of EST-SSRs is usually explained by the location of the genic SSRs in transcribed  
299 and hence conserved regions of the genome. Nevertheless, comparative studies using both types of markers  
300 showed an equivalent level of polymorphism between them, as reported in kiwifruit by Fraser et al. (2004).  
301 However, data indicated that our 20 EST-SSR markers were reasonably polymorphic in hazelnut and might  
302 be useful for germplasm characterization, and for genetic and comparative mapping.

303

#### 304 *Selection of EST-SSR markers for genetic mapping*

305 The usefulness of 36 EST-SSRs for genetic mapping in *C. avellana* was evaluated on two parents  
306 ('TGL' and 'Merveille de Bollwiller') of a  $F_1$  full-sib progeny (Table 3). This population segregates for  
307 several quantitative trait loci (QTL), such as for vigor, big bud mite (*Phytoptus avellanae* Nal.) and hazelnut  
308 weevil (*Curculio nucum* L.) tolerance, dates of budburst, flowering and nut maturity, and nut morphological  
309 traits (Beltramo et al. 2012). Pairwise comparison of 'TGL' and 'Merveille de Bollwiller' showed that 21  
310 loci were monomorphic and both parents shared the same allele. These EST-SSRs cannot be mapped in this  
311  $F_1$  population. On the other hand, 15 loci (41.6%) were heterozygous in at least one parent with either two

312 (9), three (5), and four (1) alleles. These 15 could be mapped after testing for segregation distortion in the  
313 progeny ‘TGL’ x ‘Merveille de Bollwiller’. Our results agree with similar works in other plant species.  
314 About 40% of the 87 EST-SSR developed by Chen et al. (2006) from *Citrus* spp. could be mapped in an F<sub>1</sub>  
315 population. In grape, about 52% and 35% of the loci characterized by Huang et al. (2011) could be mapped  
316 in two populations obtained from ‘Riesling’ x ‘Cabernet Sauvignon’ (*Vitis vinifera* L.) and ‘Summit’ x  
317 ‘Noble’ (*Vitis rotundifolia* Michx.), respectively.

318 EST-SSR markers are one class of marker that can contribute to ‘direct allele selection’, if they are  
319 shown to be completely associated or even responsible for a targeted trait (Varshney et al. 2005). For  
320 example, two homolog genes BpMADS2 (a homeotic B-function gene for the specification of the identity  
321 of petals and stamens) and the major pollen allergen *BetV 1* (homolog of the hazelnut pollen allergen Cor a  
322 1.04) were mapped in hazelnut using the EST-SSR markers named AJ490266 and Z72433, respectively,  
323 both developed from *Betula* spp. (Gürcan and Mehlenbacher 2010a). The BLASTx analysis of the 15  
324 selected EST-SSRs identified significant homology for 13 loci (Table 3) and 11 of them were functionally  
325 annotated using the Blast2GO program (Online data 5). In GO terms of biological process, three EST-SSRs  
326 (AG4314, BP0585, and Corav2241) were associated with biotic and abiotic stresses. AG4314 (zinc finger  
327 A20 and AN1 domain-containing stress-associated protein 5-like) was classified into the subcategory named  
328 “response to water deprivation”. A20/AN1 zinc-finger domain containing stress-associated proteins (SAPs)  
329 are considered important candidates to impart abiotic stress tolerance in plants to help protect them against  
330 severe yield loss. The first plant A20/AN1 protein identified (OsSAP1) showed multiple stress  
331 responsiveness and could confer salt, cold and dehydration stress tolerance in transgenic tobacco  
332 (Mukhopadhyay et al. 2004). BP0585 (proteasome subunit alpha type-2-a-like) was involved also with the  
333 subcategory “defense response to bacterium”. Proteasomes are large multisubunit, multicatalytic proteases  
334 responsible for the degradation of unneeded or damaged proteins by proteolysis. Proteasomes were shown  
335 to be involved in the defense responses in different plant species (Kurepa and Smalle, 2008). Plants can  
336 sense pathogen invasions by detecting pathogen-specific oligosaccharides, proteins or lipids. Also known  
337 as elicitors, proteasomes are thought to interact with receptors expressed by the plant cell, leading to the

338 activation of defence responses. Corav2241 (MACPF domain-containing protein cad1-like) was associated  
339 with several subcategories involved in defense responses. Plants respond to pathogen infection by activating  
340 a defense mechanism known as plant immunity. One of the most efficient and immediate resistance  
341 reactions against pathogen attack in plants is the hypersensitive response (HR), which leads to rapid local  
342 cell death at the site of pathogen entry and is characterized by the restricted growth and spread of the  
343 pathogen. In many cases, resistance is associated with increased expression of defense genes, including the  
344 pathogenesis-related (PR) genes and accumulation of salicylic acid (SA) in the infected leaves. SA has  
345 emerged as a key signaling component that activates both hypersensitive response (HR) and PR gene  
346 expression (Heath 2000). The CAD1 gene encodes a protein containing a domain with significant homology  
347 to the MACPF (membrane attack complex and perforin) domain of complement components and perforin  
348 proteins that are involved in innate immunity in animals (Tsutsui et al. 2008). EC numbers were reported  
349 for loci Corav1859 (reductase), Corav2208 (kinase C), and Corav2564 (pectin depolymerase) that are  
350 putatively involved in four KEGG pathways: “oxidative phosphorylation” for Corav1859,  
351 “phosphatidylinositol signaling system” for Corav2208, “starch and sucrose metabolism” and “pentose and  
352 glucuronate interconversions” for Corav2564.

353

#### 354 *Cross-species transferability within Corylus genus*

355 Transferability of the 36 primer pairs that amplified *C. avellana* and the species from which they were  
356 isolated, was evaluated in one or two samples of nine *Corylus* species (Table 4). As the primers were  
357 designed from genic regions, they were expected to amplify in related species.

358 The cross-species amplification was successful when either one or two distinct fluorescent peaks were  
359 observed after capillary electrophoresis. The success rate of markers transferability (excluding *C. avellana*)  
360 ranged from 11% (Bpt4399 and Corav6822) to 100% (Bpt5301) with an average of 55%. Bpt4399 and  
361 Corav6822 amplified only in *C. fargesii* and *C. americana*, respectively. Only two primers pairs (Ag7454  
362 and Ag8485) failed to amplify all accessions. The 12 EST-SSRs from *Alnus* amplified 0% to 78% (average  
363 50.9%) of the accessions, while 12 primer pairs developed from *Betula* amplified 11% to 100% (average

364 55.6%). Similar results were reported for 10 EST-SSR loci developed from *Betula* by Gurcan and  
365 Mehlenbacher (2010a). They amplified 70% to 100% of the *Betula* accessions and 0% to 86% of the  
366 accessions from other Betulaceae genera. Amplification of 12 EST-SSR loci from hazelnut transcriptome  
367 ranged from 11% to 89% (average 57.4%) and was very low in comparison to that reported among species  
368 of the same genera in similar works (Feng et al. 2009; Wöhrmann and Weising 2011; Acuña et al. 2012). In  
369 six out of nine *Corylus* species, the transferability rates of loci ranged from 63.9% for *C. ferox* to 88.9% for  
370 *C. americana* and this range agrees with those reported in other studies. In *Ananas* spp. (Wöhrmann and  
371 Weising 2011) and in *Eucalyptus* spp. (Acuña et al. 2012), the cross-species transferability ranged from  
372 88.9% to 100% and from 70% to 79%, respectively. On the contrary, very low values were observed in *C.*  
373 *cornuta* (5.6%), *C. colurna* (11.1%), and *C. sieboldiana* (11.1%).

374 In general, the cross-species transferability of our EST-SSRs among *Corylus* spp. was low and these  
375 data were unexpected, since the transferability of the genic SSRs in related species or genera is generally  
376 high, particularly in comparison to the gSSRs, because they originate from more conserved coding regions.  
377 The cross-amplification of gSSRs is usually restricted to congeners, whereas EST-SSRs are frequently  
378 transportable across genera and sometimes also across subfamilies or even families. Nevertheless, the  
379 amplification rates across *Corylus* spp. of gSSR loci isolated in *C. avellana* were high, in comparison to that  
380 observed with our EST-SSRs. Bassil et al. (2013) tested 23 gSSRs in 114 *Corylus* accessions representing  
381 11 species, obtaining a cross-species transferability of 74-100%. In Gürcan and Mehlenbacher (2010a) the  
382 percentage of *Corylus* accessions that amplified with two sets of 75 and 147 gSSR loci was 46.2-100% and  
383 50-100%, respectively. However, 16 EST-SSR loci characterized in this work were particularly well  
384 conserved and showed consistent amplification in more than 60% of the nine *Corylus* species analyzed.

385

## 386 **Conclusions**

387 A SSR survey from Betulaceae dbEST is reported in this study. A total of 1,282 non-redundant ESTs  
388 contained SSRs and were annotated with GO terms. This database, together with the hazelnut transcriptome,  
389 represents an important gene pool for Betulaceae improvement and a valuable resource for developing PCR-



390 based genetic markers, in particular for the European hazelnut. Markers developed from *Alnus* and *Betula*  
391 showed a relatively high level of transferability (69.3%) in these species.

392 The EST-SSRs developed in this study increase the number of SSR markers currently available for  
393 *C. avellana*. Many showed good amplification and polymorphism in hazelnut and are suitable marker  
394 candidates for genotyping and population genetic analyses. Among them, a set of EST-SSRs was selected  
395 as promising for genetic mapping and marker-assisted selection in hazelnut. Moreover, when mapped, they  
396 can be used in synteny studies among *Corylus* species to better understand interspecific gene flow, genome  
397 organization, and evolutionary relationships.

398

399

400

401 **Acknowledgements** The research was funded by the Foundation Cassa di Risparmio di Torino (CRT).

402

403 **References**

- 404 Acuña CV, Fernandez P, Villalba PV, García MN, Hopp HE, Marcucci Poltri SN (2012) Discovery, validation, and in  
405 silico functional characterization of EST-SSR markers in *Eucalyptus globulus*. *Tree Genet Genomes* 8: 289-301
- 406 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology.  
407 The Gene Ontology Consortium. *Nat Genet* 25: 25–29
- 408 Bacchetta L, Aramini M, Zini A, Di Giammatteo V, Spera D, Drogoudi P, Rovira M, Silva AP, Solar A, Botta R (2013)  
409 Fatty acids and alpha-tocopherol composition in hazelnut (*Corylus avellana* L.): a chemometric approach to  
410 emphasize the quality of European germplasm. *Euphytica* 191: 57-73
- 411 Bassil NV, Botta R, Mehlenbacher SA (2005a) Microsatellite markers in the hazelnut: isolation, characterization and  
412 cross-species amplification in *Corylus*. *J Am Soc Hort Sci* 130: 543-549
- 413 Bassil NV, Botta R, Mehlenbacher SA (2005b) Additional microsatellites of the European hazelnut. *Acta Hort* 686:  
414 105-110
- 415 Bassil NV, Boccacci P, Botta R, Postman J, Mehlenbacher SA (2013) Nuclear and chloroplast microsatellite markers  
416 to assess genetic diversity and evolution in hazelnut species, hybrids and cultivars. *Genet Resour Crop Evol* 60:  
417 543–568
- 418 Beltramo C, Boccacci P, Sandoval Prando MA, Portis E, Botta R (2012) Development of a genetic linkage map in  
419 hazelnut (*Corylus avellana* L.) for the detection of QTLs. VIII International Congress on Hazelnut, Temuco City  
420 (Chile), March 19-22. *Acta Hort*: in press
- 421 Boccacci P, Botta R (2010) Microsatellite variability and genetic structure in hazelnut (*Corylus avellana* L.) cultivars  
422 from different growing regions. *Sci Hort* 124: 128-133
- 423 Boccacci P, Akkak A, Bassil NV, Mehlenbacher SA, Botta R (2005) Characterization and evaluation of microsatellite  
424 loci in European hazelnut (*Corylus avellana* L.) and their transferability to other *Corylus* species. *Mol Ecol Notes*  
425 5: 934-937
- 426 Boccacci P, Akkak A, Botta R (2006) DNA-typing and genetic relationships among European hazelnut (*Corylus*  
427 *avellana* L.) cultivars using microsatellite markers. *Genome* 49: 598-611
- 428 Boccacci P, Rovira M, Botta R (2008) Genetic diversity of hazelnut (*Corylus avellana* L.) germplasm in northeastern  
429 Spain. *HortScience* 43: 667-672

430 Boccacci P, Aramini M, Valentini N, Bacchetta L, Rovira M, Drogoudi P, Silva AP, Solar A, Calizzano F, Erdoğan  
431 V, Cristofori V, Ciarmiello LF, Contessa C, Ferreira JJ, Marra FP, Botta R (2013). Molecular and morphological  
432 diversity of on-farm hazelnut (*Corylus avellana* L.) landraces from southern Europe and their role in the origin  
433 and diffusion of cultivated germplasm. *Tree Genet Genomes* 9: 1465–1480

434 Botstein D, White RL, Skolnick M and Davis RW (1980) Construction of a genetic linkage map in man using  
435 restriction fragment length polymorphisms. *Am J Hum Genet* 32: 314-331

436 Campa A, Trabanco E, Pérez-Vega E, Rovira M, Ferreira JJ (2011) Genetic relationship between cultivated and wild  
437 hazelnuts (*Corylus avellana* L.) collected in northern Spain. *Plant Breed* 130: 360-366

438 Chen ZD, SR Manchester, HY Sun (1999) Phylogeny and evolution of the Betulaceae as inferred from DNA  
439 sequences, morphology, and paleobotany. *Am J Bot* 86: 1168–1181

440 Chen C, Zhou P, Choi YA, Huang S, Gmitter Jr FG (2006) Mining and characterizing microsatellites from *Citrus*  
441 ESTs. *Theor Appl Genet* 112: 1248–1257

442 Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation,  
443 visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676

444 Decroocq V, Fave MG, Hagen L, Bordenave L, Decroocq S (2003). Development and transferability of apricot and  
445 grape EST microsatellite markers across taxa. *Theor Appl Genet* 106: 912-922

446 Edwards KJ, Barker JHA, Daly A, Jones C, Karp A (1996) Microsatellite libraries enriched for several microsatellite  
447 sequences in plants. *BioTechniques* 20: 758–760

448 Erdoğan V, Mehlenbacher SA (2000) Phylogenetic relationships of *Corylus* species (Betulaceae) based on nuclear  
449 ribosomal DNA ITS region and chloroplast matK gene sequences. *Syst Bo* 25:727–727

450 FAOstat (2014) Agriculture data. Available from: <http://faostat.fao.org/site/339/default.aspx>. Accessed 26 Jnuary 2014

451 Feng SP, Li WG, Huang HS, Wang JY, Wu YT (2009) Development, characterization and cross-species/genera  
452 transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). *Mol Breeding* 23: 85–97

453 Fraser LG, Harvey CF, Crowhurst RN, De Silva HN (2004) ESTderived microsatellites from Actinidia species and  
454 their potential for mapping. *Theor Appl Genet* 108: 1010–1016

455 Gasic K, Han Y, Kertbundit S, Shulaev V, Iezzoni AF, Stover EW, Bell RL, Wisniewski ME, Korban SS (2009)  
456 Characteristics and transferability of new apple EST-derived SSRs to other Rosaceae species. *Mol Breeding* 23:  
457 397–411

458 Gökirmak T, Mehlenbacher SA, Bassil NV (2009) Characterization of European hazelnut (*Corylus avellana* L.)  
459 cultivars using SSR markers. *Genet Resour Crop Evol* 56: 147-172

460 Gürcan K, Mehlenbacher SA (2010a) Transferability of microsatellite markers in the Betulaceae. *J Amer Soc Hort Sci*  
461 135: 159-173

462 Gürcan K, Mehlenbacher SA (2010b) Development of microsatellite marker loci for European hazelnut (*Corylus*  
463 *avellana* L.) from ISSR fragments. *Mol Breed* 26: 551-559

464 Gürcan K, Mehlenbacher SA, Botta R, Boccacci P (2010a) Development, characterization, segregation, and mapping  
465 of microsatellite markers for European hazelnut (*Corylus avellana* L.) from enriched genomic libraries and  
466 usefulness in genetic diversity studies. *Tree Genet Genomes* 6: 513-531

467 Gürcan K, Mehlenbacher SA, Erdoğan V (2010b) Genetic diversity in hazelnut (*Corylus avellana* L.) cultivars from  
468 Black Sea countries assessed using SSR markers. *Plant Breed* 129: 422-434

469 Heath MC (2000) Hypersensitive response-related death, *Plant Mol. Biol.* 44: 321–334

470 Heesacker A, Kishore VK, Gao W, Tang S, Kolkman JM, Gingle A, Matvienko M, Kozik A, Michelmore RM, Lai Z,  
471 Rieseberg LH, Knapp SJ (2008) SSRs and INDELS mined from the sunflower EST database: abundance,  
472 polymorphisms, and cross-taxa utility. *Theor Appl Genet* 117: 1021–1029

473 Huang H, Lu J, Hunter W, Dowd S, Dang P (2011). Mining and validating grape (*Vitis* L.) ESTs to develop EST-SSR  
474 markers for genotyping and mapping. *Mol Breed* 28: 241-254

475 Järvinen, P, Palmé A, Morales LO, Länneppää M, Keinänen M, Sapanen T, Lascoux M (2004). Phylogenetic  
476 relationships of *Betula* species (Betulaceae) based on nuclear ADH and chloroplast matK sequences. *Amer J Bot*  
477 91: 1834–1845

478 Kurepa J, Smalle JA (2008) Structure, function and regulation of plant proteasomes. *Biochimie* 90: 324-335

479 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A,  
480 Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*  
481 23: 2947-2948

482 Liu K, Muse SV (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*  
483 21: 2128-2129

484 Mehlenbacher SA, Brown RN, Nouhra EN, Gökirmak T, Bassil NV, Kubisiak TL (2006) A genetic linkage map for  
485 hazelnut (*Corylus avellana* L.) based on RAPD and SSR markers. *Genome* 49: 122–133

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

488 Mukhopadhyay A, Vij S, Tyagi AK (2004). Overexpression of a zinc-finger protein gene from rice confers tolerance  
489 to cold, dehydration, and salt stress in transgenic tobacco. *Proc Natl Acad Sci USA* 101: 6309–6314

490 Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci* 1:  
491 215–222

492 Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol*  
493 *Biol* 132: 365-386

494 Rowley ER, Fox SE, Bryant DW, Sullivan CM, Priest HD, Givan SA, Mehlenbacher SA, Mockler TC (2012)  
495 Assembly and characterization of the European hazelnut ‘Jefferson’ transcriptome. *Cro Sci* 52: 2679-2686

496 Sathuvalli VR, Mehlenbacher SA (2012) Characterization of American hazelnut (*Corylus americana*) accessions and  
497 *Corylus americana* × *Corylus avellana* hybrids using microsatellite markers. *Genet Resour Crop Evol* 59: 1055-  
498 1075

499 Sathuvalli VR, Chen H, Mehlenbacher SA, Smith DC (2011) DNA markers linked to eastern filbert blight resistance  
500 in ‘Ratoli’ hazelnut (*Corylus avellana* L.). *Tree Genet Genomes* 7: 337–345

501 Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and  
502 characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 106: 411–  
503 422

504 Thomas MR, Matsumoto S, Cain P, Scott NS (1993) Repetitive DNA of grapevine: classes present and sequences  
505 suitable for cultivar identification. *Theor Appl Genet* 86: 173-180

506 Tsutsui T, Asada Y, Tamaoki M, Ikeda A, Yamaguchi J (2008) Arabidopsis CAD1 negatively controls plant immunity  
507 mediated by both salicylic acid-dependent and -independent signalling pathways. *Plant Science* 175: 604-611

508 Varshney R, Graner A, Sorrells M (2005) Genic microsatellite markers in plants: features and applications. *Trends*  
509 *Biotech* 23: 48–55

510 Victoria FC, da Maia LC, Costa de Oliveira A (2011) *In silico* comparative analysis of SSR markers in plants. *BMC*  
511 *Plant Biology* 11: 15

512 Whitcher IN, Wen J (2001) Phylogeny and biogeography of *Corylus* (Betulaceae): inference from ITS sequences. *Syst*  
513 *Bot* 26: 283–298

514 Wöhrmann T, Weising K (2011) In silico mining for simple sequence repeat loci in a pineapple expressed sequence  
515 tag database and cross-species amplification of EST-SSR markers across Bromeliaceae. Theor Appl Genet 123:  
516 635–647

517 Yoo K-O, Wen J (2002) Phylogeny and biogeography of *Carpinus* and subfamily *Coryloideae* (Betulaceae). Int J Pl  
518 Sci 163: 641–650

519 Zhang LS, Becquet V, Li SH, Zhang D (2003) Optimization of multiplex PCR and multiplex gel electrophoresis in  
520 sunflower SSR analysis using infrared fluorescence and tailed primers. Acta Botanica Sinica 45: 1312-1318

521

522

523

524

525

526 **List of figures:**

527 **Figure 1** - GO assignment (second level GO terms) of 921 non-redundant EST-SSR sequences from  
528 Betulaceae. The total numbers of EST-SSRs annotated for each main category were: 560 for biological  
529 process, 509 for molecular function, and 494 for cellular component

530

531 **Table 1** – Characteristics of 36 validates primer pairs derived from *Alnus glutinosa* (Ag), *Betula pendula* (Bp), *Betula plathyphylla* (Bpt), and *Corylus*  
 532 *avellana* (Corav) SSR-containing EST sequences

Primer ID	Accession number	Repeat type	Forward primer (5' - 3')	Reverse primer (5' - 3')	Annealing (°C)		Expected size (bp)	Allele size range (bp)
					step I	step II		
Ag0100	FQ359611	(CT) <sub>10</sub>	GGATCTCTGTGGCATTGTT	TCTGGATTCTGGAAGGCACT	60	54	151	134-170
Ag0851	FQ356617	(TC) <sub>15</sub>	CATGCCGGACAATAACAACAC	GCACAGCACCAACACAAAAC	60	54	156	162-197
Ag2912	FQ357710	(CT) <sub>10</sub>	TCGCTCTCTCAAACCCAAAT	TATCACTATCGCGGTCAGCA	58	54	165	158-186
Ag3023	FQ357821	(CT) <sub>7</sub>	CGATAACAACGCTACCGACAG	ACGTGCGAGATACGCTGTAA	60	54	198	209-228
Ag3754	FQ355542	(CT) <sub>8</sub>	GGGTTTATGAATCGCCAGAA	GTTAGCAGAGCATGGTGCAG	60	54	158	209-195
Ag3765	FQ338535	(CTT) <sub>5</sub>	GCTGGGCTTATTGGTGGTAA	TGGGTTGGATCCTCAAAGTC	56	53	151	170-173
Ag4306	FQ360650	(CT) <sub>9</sub>	GCCACGGTAATCACAGACT	AAAACAGGCGGAAAACAACA	56	53	221	202-241
Ag4314	FQ360658	(TC) <sub>10</sub>	CCGCTCAATCCCTTCAAAT	GGTATGGTTTCGGGGACTTT	56	53	217	215-241
Ag4395	FQ360739	(CT) <sub>12</sub>	TACGGGGAAGAGGAGGTAGG	TGTCCCAGAAGTTTGCCTCT	56	53	153	167-201
Ag4765	FQ353474	(CT) <sub>11</sub>	GGGAGCTAAAACAGCCAAAA	CCCATCAGGCAGGTAAACAA	56	53	168	171-190
Ag7454	FQ334036	(CTT) <sub>5</sub>	GACGGGATTTGGAAGATTGA	AATGAAGCGATTTGGCAGTC	56	53	162	181-229
Ag8485	FQ331021	(CTT) <sub>4</sub>	AGTCGGTGGCAGAGAAGAAG	TCTCGGGCTTCTTTGTTGTT	56	53	151	168-171
Bp0326	CD278280	(GAA) <sub>3</sub>	CACCAACAGACCATGGAAAA	CACCCGGAGAAGTTCAAGAG	58	54	202	208-221
Bp0585	CD278539	(CT) <sub>9</sub>	TTGACATTTTCTCTCCTTGACG	TCAAAGCGTGTTCAATCTGC	58	54	157	273-293
Bp8823	CD276777	(CTT) <sub>5</sub>	CAAAGCAATGCTCTTACCA	AATCGAGCACAACTGCTCCT	60	54	159	172-187
Bp8953	CD276907	(CT) <sub>9</sub>	GGTTGCTCAACCTAACCAACA	TGGACAAGAACAACCACCAA	60	54	221	226-252
Bp9171	CD277125	(CT) <sub>11</sub>	AACCAACCAGCCAAGTGAAC	CTGCACCTCCCAACAGTTCT	56	53	229	246-267
Bpt1648	FG067245	(CT) <sub>11</sub>	AAATCCTCATAGTCCGCTCAA	GAGGCGGAGAAACAGAACAG	60	54	193	165-211
Bpt2145	FG065565	(GGT) <sub>6</sub>	GTAATTCAGGTGCCCACTCG	ACTGAGGTGGTCGGAATCAA	56	53	198	210-229
Bpt3751	FG067049	(CT) <sub>10</sub>	GGGAGAGCATTTTACAGAGTGG	AATGGTCCATAGCCCAGCTT	56	53	184	198-207
Bpt4399	FG067057	(GGT) <sub>5</sub>	TATCTGGTGCTGCAATTTGG	AGAGGCATGGTCTCTGTTC	58	54	186	193-206
Bpt5301	FG065983	(CTT) <sub>5</sub>	CAGACCCAGAAACCCAGAGA	CGGACTCGTTATCGTCCACT	60	54	169	184-190
Bpt5452	FG067556	(TA) <sub>10</sub>	TTAGCTGTGCTGCGAATGAC	ATAACATGCGGACGTCAGTG	58	54	168	161-194
Bpt9403	FG066704	(TGG) <sub>6</sub>	CGCTCCATACGAAGAACGA	GGAGCTTTGGATCTGAGGTG	56	53	150	165-169
Corav7591	Corav7591	(TA) <sub>8</sub>	AGGGAATTCAGATGCCAAAA	CTCCTCTTACAGGCTCCAGTG	60	54	213	216-249

Corav1232	Corav1232	(CT) <sub>8</sub>	CCTTCACCGTTACACCCTCT	CACCAGAGAAATTCCCAACG	58	54	175	184-204
Corav1576	Corav1576	(GA) <sub>11</sub>	CAGCCACTTCAGCTCACAAA	ACTTTCCAAGATTGCGGATT	58	54	173	186-202
Corav1859	Corav1859	(GA) <sub>9</sub>	ACCTCCATGCCAGAGATGAT	CATTTTCTGGCCCCTTCTTT	56	53	188	202-219
Corav2208	Corav2208	(GA) <sub>10</sub>	TGTGTTCCATTGGTCTCAGC	CATTGGAAGAACTCCCCTGA	56	53	184	201-237
Corav2241	Corav2241	(CT) <sub>8</sub>	ACCCACGAGAAATTGGAGTG	GGGGAAGTGGGGATAGAAAG	58	54	231	238-253
Corav2387	Corav2387	(CT) <sub>8</sub>	AGGGTTCACAGTGATGACAC	TTCTTCGATCCCTCTTCTCA	56	53	216	225-239
Corav2560	Corav2560	(CT) <sub>11</sub>	TCCTTCTCCTCTCCCTCCTC	CACACGGTAACAAAGGCAAA	58	54	166	177-197
Corav2564	Corav2564	(CT) <sub>9</sub>	CCCAGTCCCCCTTATAAACC	TGGGTGGAGATTTTGGAAAG	60	54	171	179-191
Corav4911	Corav4911	(TA) <sub>8</sub>	CGCTTGTATGTCACCTTCC	GCCTACGAAAAGATCGCTTG	56	53	177	178-204
Corav6822	Corav6822	(TA) <sub>9</sub>	ATGGGTGAGAGAGACCTGGA	AGCACTTTGAAACACCACCA	58	54	164	179-189
Corav692	Corav692	(AG) <sub>9</sub>	CCAAGCCTAGAGCGAGAGAG	CGCACTGTCAGATCTCTCCA	58	54	154	166-182

533

534



535 **Table 2** – Number of genotypes ( $N_G$ ), number of alleles ( $N_A$ ), gene diversity ( $H_e$ ), observed heterozygosity  
 536 ( $H_o$ ), and polymorphic information content (PIC) of 20 polymorphic EST-SSR markers in 18 *C. avellana*  
 537 cultivars.

538

<b>Marker</b>	<b><math>N_G</math></b>	<b><math>N_A</math></b>	<b><math>H_e</math></b>	<b><math>H_o</math></b>	<b>PIC</b>
Ag0100	5	5	0.56	0.17	0.48
Ag0851	9	6	0.73	0.39	0.68
Ag2912	6	4	0.59	0.56	0.51
Ag3754	4	3	0.52	0.56	0.41
Ag4314	3	4	0.25	0.22	0.24
Ag4395	10	8	0.80	0.83	0.77
Bp0585	9	6	0.71	0.50	0.66
Bpt5452	6	5	0.63	0.78	0.57
Corav692	2	2	0.24	0.28	0.21
Corav1232	9	7	0.80	0.61	0.77
Corav1576	9	5	0.75	0.89	0.71
Corav1859	8	6	0.70	0.72	0.66
Corav2208	5	5	0.57	0.89	0.49
Corav2241	2	2	0.15	0.17	0.14
Corav2387	4	3	0.43	0.50	0.39
Corav2560	14	8	0.83	0.89	0.81
Corav2564	3	2	0.50	0.56	0.38
Corav4911	3	5	0.58	1.00	0.49
Corav6822	4	3	0.51	0.56	0.43
Corav7591	3	3	0.33	0.28	0.30
<b>Mean</b>	5.90	4.60	0.56	0.57	0.50

539

540

541 **Table 3** - Distribution of the segregation types expected in the F<sub>1</sub> population ‘TGL’ x ‘Merveille de Bollwiller’ (syn ‘Hall’s Giant’) and putative  
542 function of 15 EST-SSR markers selected for hazelnut mapping (NA= not assigned). Expected ratio is assumed that there are not null alleles. F and  
543 M indicates markers that segregate from the female and male parents, respectively

544

Marker	TGL	Merveille de Bollwiller	Alleles	Expected ratio	Putative function	min e-value	sim mean (%)
AG2912	169/169	169/171	2	1:1 M	predicted protein	7.7 e-4	72.0
AG3754	170/170	170/172	2	1:1 M	NA	-	-
AG4314	233/235	233/233	2	1:1 F	zinc finger A20 and AN1 domain-containing stress-associated protein 5-like	2.9 e-31	68.1
AG4395	180/182	182/184	3	1:1:1:1	NA	-	-
BP0585	281/287	281/283	3	1:1:1:1	proteasome subunit alpha type-2-a-like	1.1 e-117	97.6
BPT5452	167/188	167/179	3	1:1:1:1	isoflavone reductase family protein	4.0 e-6	78.5
Corav1232	184/201	184/198	3	1:1:1:1	udp-galactose transporter 1-like	0.0 e0	94.0
Corav1576	192/198	192/196	3	1:1:1:1	ring finger and transmembrane domain-containing protein 2-like	0.0 e0	78.8
Corav1859	206/209	209/209	2	1:1 F	ubiquinol-cytochrome c reductase iron-sulfur subunit family protein	1.5 e-157	84.4
Corav2208	203/205	203/205	2	1:2:1	3-phosphoinositide-dependent protein kinase 2-like	0.0 e0	91.8
Corav2241	248/248	248/253	2	1:1 M	MACPF domain-containing protein cad1-like	0.0 e0	86.3
Corav2560	181/189	179/185	4	1:1:1:1	<i>bes1 bzr1</i> homolog protein 2-like	9.2 e-77	88.7
Corav2564	181/187	187/187	2	1:1 F	probable polygalacturonase-like	0.0 e0	89.8
Corav4911	185/197	185/197	2	1:2:1	ribosomal protein s18	2.6 e-59	92.4
Corav6822	183/185	183/183	2	1:1 F	mediator of RNA polymerase II transcription subunit 33a-like	3.7 e-69	86.7

545

**Table 4** – Cross-species amplification of 36 EST-SSR markers among 9 *Corylus* species.

Marker	<i>C. americana</i> (n = 2)	<i>C. cali formica</i> (n = 2)	<i>C. chinensis</i> (n = 2)	<i>C. colurna</i> (n = 2)	<i>C. cornuta</i> (n = 1)	<i>C. heterophylla</i> (n = 1)	<i>C. fargesii</i> (n = 1)	<i>C. ferox</i> (n = 1)	<i>C. sieboldiana</i> (n = 2)	Total (%)
Ag0100	+	+	+	-	-	+	+	+	-	67
Ag0851	+	+	-	-	-	+	+	+	-	56
Ag2912	+	-	+	-	-	+	+	+	-	56
Ag3023	+	+	+	-	-	+	+	+	-	67
Ag3754	+	+	+	-	-	+	+	+	-	67
Ag3765	+	-	+	-	-	-	+	-	-	33
Ag4306	+	+	+	-	-	+	+	-	-	56
Ag4314	+	+	+	-	-	+	+	+	-	67
Ag4395	+	+	+	-	-	+	+	+	-	67
Ag4765	+	+	+	-	-	+	+	+	+	78
Ag7454	-	-	-	-	-	-	-	-	-	0
Ag8485	-	-	-	-	-	-	-	-	-	0
Bp0326	+	+	+	+	-	+	+	+	-	78
Bp0585	+	-	+	-	-	-	+	+	-	44
Bp8823	+	+	+	-	-	+	+	+	+	78
Bp8953	+	+	+	-	-	-	+	-	-	44
Bp9171	+	+	+	-	-	-	+	-	-	44
Bpt1648	+	-	+	-	-	-	+	-	-	33
Bpt2145	+	+	+	-	-	+	+	+	-	67
Bpt3751	+	+	+	-	-	-	+	-	-	44
Bpt4399	-	-	-	-	-	-	+	-	-	11
Bpt5301	+	+	+	+	+	+	+	+	+	100
Bpt5452	+	+	+	-	-	+	-	+	-	67
Bpt9403	+	+	+	-	-	+	+	-	-	67
Corav692	+	+	+	-	-	+	+	+	-	67
Corav1232	-	+	+	-	-	+	+	+	-	67
Corav1576	+	+	+	-	-	+	+	+	-	56
Corav1859	+	+	-	-	-	+	-	+	-	44
Corav2208	+	+	+	+	-	+	+	+	-	78
Corav2241	+	-	-	-	-	+	-	+	-	33
Corav2387	+	+	+	-	-	+	+	+	-	67
Corav2560	+	+	+	-	-	+	+	+	+	78
Corav2564	+	+	+	+	+	+	+	+	-	89
Corav4911	+	+	+	-	-	-	+	-	-	44
Corav6822	+	-	-	-	-	-	-	-	-	11
Corav7591	+	+	+	-	-	+	+	-	-	56
<b>Total (%)</b>	88.9	75	80.6	11.1	5.6	69.4	83.3	63.9	11.1	