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

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

Boccacci P.[1,2,*)], Sandoval Prando M.A.[2,*)], Beltramo C. [2)], Lembo A. [3)], Sartor C. [2)], Mehlenbacher S.A. [4)], Botta R. [2)], Torello Marinoni D. [2)]

* These authors contributed equally to this work

[1)] Institute for Sustainable Plant Protection - National Research Council (IPSP-CNR), Unit of Grugliasco, Largo Paolo Braccini, 2 - 10095 Grugliasco (TO), Italy

[2)] Department of Agriculturae, Forestry and Food Science (DISAFA) - University of Turin, Largo Paolo Braccini, 2 - 10095 Grugliasco (TO), Italy

[3)] Molecular Biotechnology Center, Department of Molecular Biotechnology and Health Sciences, University of Turin, Via Nizza 52 – 10126 Torino, Italy

[4)] Department of Horticulture, Oregon State University, Agricultural and Life Sciences Bldg. 4017, Corvallis, OR 97331, USA

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

Abstract

The European hazelnut (*Corylus avellana* L.) is one of the most important nut crops. In this work we characterize functional microsatellite or simple sequence repeat (SSR) markers for genetic analysis and molecular breeding in this species. A total of 38,454 Betulaceae EST sequences from NCBI resulted in 1,282 non-redundant EST-SSRs. Dinucleotide repeats were the most abundant (63.9%), followed by trinucleotides (33.8%). The putative functions of the non-redundant EST-SSRs were classified according to gene ontology (GO) categories (biological process, molecular function, and cellular component). A total of 921 sequences showed significant hits with the non-redundant protein database, and GO categories were assigned to 696 (75.5%) of them. Flanking primer pairs were designed for 78 di- and trinucleotide EST-SSRs from *Alnus glutinosa* L. (29), *Betula pendula* Roth (26), and *Betula platyphylla* Suckaczev (23). Further, 41 dinucleotide repeats selected from hazelnut transcriptome sequences were added. Thirty-six out 119 primer pairs generated amplification products in six hazelnut accessions and in the samples of the species from which they were isolated. Among them, 20 were polymorphic when tested on 18 hazelnut cultivars. Fifteen loci are suitable for mapping in a F_1 population of 'Tonda Gentile delle Langhe' x 'Merveille de Bollwiller' and 11 of them were functionally annotated. The cross-species transferability of 36 EST-SSR loci within nine *Corylus* species was also performed. The success rate of markers transferability (including *C. avellana*) ranged from 11% to 100%, with an average of 55%. The EST-SSRs developed increase the number of markers currently available for hazelnut.

Key words: expressed sequence tags; microsatellite; functional annotation; polymorphism; transferability; filbert

Introduction

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

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

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

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

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

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

The present study aims to develop conserved orthologous markers for genetic analysis of different Betulaceae species and to expand the genomic resources for the European hazelnut breeding and mapping. To achieve this, an *in silico* identification and characterization of unique SSRs derived from Betulaceae ESTs, retrieved from a public database, was performed. Microsatellites were identified in cDNA sequences of black alder (*Alnus glutinosa* L.), silver birch (*Betula pendula* Roth), and Japanese white birch (*Betula platyphylla* Suckaczev). In addition, SSRs were selected from *C. avellana* transcriptome sequences. A set of EST-SSR markers was developed and evaluated for their ability to detect polymorphism in hazelnut, suitability for genetic mapping in a F_1 full-sib progeny of 'Tonda Gentile delle Langhe' x 'Merveille de Bollwiller', and transferability across the genus *Corylus*.

Materials and methods

Mining of SSR-containing sequences and functional annotation

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

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

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

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

Plant materials and DNA extraction

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

PCR amplification and fragment analysis

Sequences flanking the microsatellite motifs in 119 ESTs were used to design primer pairs using Primer3 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 contings
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 (Tm) 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 ≥ 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 µl containing 40 ng DNA, 1x Biolase NH_4
166 reaction buffer [160 mM $(NH_4)_2SO_4$, 670 mM Tris–HCl (pH 8.8 at 25°C), 0.1% Tween-20], 2.25 mM
167 $MgCl_2$, 200 µM of each dNTPs, 0.75 µl of a primer mix containing 8 pmol of reverse primer and 2 pmol of
168 the forward primer, 1.2 µl of 5 pmol M13 tail primer, and 0.4 U Biolase DNA polymerase 5U/µl (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),

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

Results and discussion

Identification, frequency and characterization of EST-SSRs

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

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

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

Functional annotation of ESTs containing SSRs

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

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

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

EST-SSR validation and polymorphism in C. avellana

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

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

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

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

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

Selection of EST-SSR markers for genetic mapping

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

(9), three (5), and four (1) alleles. These 15 could be mapped after testing for segregation distortion in the progeny 'TGL' x 'Merveille de Bollwiller'. Our results agree with similar works in other plant species. About 40% of the 87 EST-SSR developed by Chen et al. (2006) from *Citrus* spp. could be mapped in an F_1 population. In grape, about 52% and 35% of the loci characterized by Huang et al. (2011) could be mapped in two populations obtained from 'Riesling' x 'Cabernet Sauvignon' (*Vitis vinifera* L.) and 'Summit' x 'Noble' (*Vitis rotundifolia* Michx.), respectively.

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

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

Cross-species transferability within Corylus *genus*

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

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

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

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

Conclusions

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

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

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

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

References

Acuña CV, Fernandez P, Villalba PV, García MN, Hopp HE, Marcucci Poltri SN (2012) Discovery, validation, and in silico functional characterization of EST-SSR markers in *Eucalyptus globulus*. Tree Genet Genomes 8: 289-301

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29

Bacchetta L, Aramini M, Zini A, Di Giammatteo V, Spera D, Drogoudi P, Rovira M, Silva AP, Solar A, Botta R (2013) Fatty acids and alpha-tocopherol composition in hazelnut (*Corylus avellana* L.): a chemometric approach to emphasize the quality of European germplasm. Euphytica 191: 57-73

Bassil NV, Botta R, Mehlenbacher SA (2005a) Microsatellite markers in the hazelnut: isolation, characterization and cross-species amplification in *Corylus*. J Am Soc Hort Sci 130: 543-549

Bassil NV, Botta R, Mehlenbacher SA (2005b) Additional microsatellites of the European hazelnut. Acta Hort 686: 105-110

Bassil NV, Boccacci P, Botta R, Postman J, Mehlenbacher SA (2013) Nuclear and chloroplast microsatellite markers to assess genetic diversity and evolution in hazelnut species, hybrids and cultivars. Genet Resour Crop Evol 60: 543–568

Beltramo C, Boccacci P, Sandoval Prando MA, Portis E, Botta R (2012) Development of a genetic linkage map in hazelnut (*Corylus avellana* L.) for the detection of QTLs. VIII International Congress on Hazelnut, Temuco City (Chile), March 19-22. Acta Hort: in press

Boccacci P, Botta R (2010) Microsatellite variability and genetic structure in hazelnut (*Corylus avellana* L.) cultivars from different growing regions. Sci Hortic 124: 128-133

Boccacci P, Akkak A, Bassil NV, Mehlenbacher SA, Botta R (2005) Characterization and evaluation of microsatellite loci in European hazelnut (*Corylus avellana* L.) and their transferability to other *Corylus* species. Mol Ecol Notes 5: 934-937

Boccacci P, Akkak A, Botta R (2006) DNA-typing and genetic relationships among European hazelnut (*Corylus avellana* L.) cultivars using microsatellite markers. Genome 49: 598-611

Boccacci P, Rovira M, Botta R (2008) Genetic diversity of hazelnut (*Corylus avellana* L.) germplasm in northeastern Spain. HortScience 43: 667-672

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

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

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

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

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

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

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

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

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

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

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

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

Gasic K, Han Y, Kertbundit S, Shulaev V, Iezzoni AF, Stover EW, Bell RL, Wisniewski ME, Korban SS (2009) Characteristics and transferability of new apple EST-derived SSRs to other Rosaceae species. Mol Breeding 23: 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, Lännenpää M, Keinänen M, Sopanen 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

List of figures:

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

Table 1 – Characteristics of 36 validates primer pairs derived from *Alnus glutinosa* (Ag), *Betula pendula* (Bp), *Betula plathyphylla* (Bpt), and *Corylus 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	$(CT)_{10}$	GGATCTCTGTGGGCATTGTT	TCTGGATTCTGGAAGGCACT	60	54	151	134-170
Ag0851	FQ356617	$(TC)_{15}$	CATGCCGGACAATACAACAC	GCACAGCACCAACACAAAAC	60	54	156	162-197
Ag2912	FQ357710	$(CT)_{10}$	TCGCTCTCTCAAACCCAAAT	TATCACTATCGCGGTCAGCA	58	54	165	158-186
Ag3023	FQ357821	$(CT)_{7}$	CGATACAACGCTACCGACAG	ACGTGCGAGATACGCTGTAA	60	54	198	209-228
Ag3754	FQ355542	$(CT)_{8}$	GGGTTTATGAATCGCCAGAA	GTTAGCAGAGCATGGTGCAG	60	54	158	209-195
Ag3765	FQ338535	$(CTT)_{5}$	GCTGGGCTTATTGGTGGTAA	TGGGTTGGATCCTCAAAGTC	56	53	151	170-173
Ag4306	FQ360650	$(CT)_{9}$	GCCCACGGTAATCACAGACT	AAAACAGGCGGAAAACAACA	56	53	221	202-241
Ag4314	FQ360658	$(TC)_{10}$	CCGCTCAATCCCTTCAAAT	GGTATGGTTTCGGGGACTTT	56	53	217	215-241
Ag4395	FQ360739	$(CT)_{12}$	TACGGGGAAGAGGAGGTAGG	TGTCCCAGAAGTTTGCCTCT	56	53	153	167-201
Ag4765	FQ353474	$(CT)_{11}$	GGGAGCTAAAACAGCCAAAA	CCCATCAGGCAGGTAAACAA	56	53	168	171-190
Ag7454	FQ334036	$(CTT)_{5}$	GACGGGATTTGGAAGATTGA	AATGAAGCGATTTGGCAGTC	56	53	162	181-229
Ag8485	FQ331021	$(CTT)_{4}$	AGTCGGTGGCAGAGAAGAAG	TCTCGGGCTTCTTTGTTGTT	56	53	151	168-171
Bp0326	CD278280	$(GAA)_{3}$	CACCAACAGACCATGGAAAA	CACCCGGAGAAGTTCAAGAG	58	54	202	208-221
Bp0585	CD278539	$(CT)_{9}$	TTGACATTTTCTCTCCTTGACG	TCAAAGCGTGTTCAATCTGC	58	54	157	273-293
Bp8823	CD276777	$(CTT)_{5}$	CAAAGCAATGCTCTTCACCA	AATCGAGCACAACTGCTCCT	60	54	159	172-187
Bp8953	CD276907	$(CT)_{9}$	GGTTGCTCAACCTAACCAACA	TGGACAAGAACAACCACCAA	60	54	221	226-252
Bp9171	CD277125	$(CT)_{11}$	AACCAACCAGCCAAGTGAAC	CTGCACCTCCCAACAGTTCT	56	53	229	246-267
Bpt1648	FG067245	$(CT)_{11}$	AAATCCTCATAGTCCGCTCAA	GAGGCGGAGAAACAGAACAG	60	54	193	165-211
Bpt2145	FG065565	$(GGT)_{6}$	GTAATTCAGGTGCCCACTCG	ACTGAGGTGGTCGGAATCAA	56	53	198	210-229
Bpt3751	FG067049	$(CT)_{10}$	GGGAGAGCATTTCAGAGTGG	AATGGTCCATAGCCCAGCTT	56	53	184	198-207
Bpt4399	FG067057	$(GGT)_{5}$	TATCTGGTGCTGCAATTTGG	AGAGGCATGGTCTCTGTTCC	58	54	186	193-206
Bpt5301	FG065983	$(CTT)_{5}$	CAGACCCAGAAACCCAGAGA	CGGACTCGTTATCGTCCACT	60	54	169	184-190
Bpt5452	FG067556	$(TA)_{10}$	TTAGCTGTGCTGCGAATGAC	ATAACATGCGGACGTCAGTG	58	54	168	161-194
Bpt9403	FG066704	$(TGG)_{6}$	CGCTCCATACGAAGAACGA	GGAGCTTTGGATCTGAGGTG	56	53	150	165-169
Corav7591	Corav7591	$(TA)_{8}$	AGGGAATTCAGATGCCAAAA	CTCCTCTTCAGGCTCCAGTG	60	54	213	216-249

Corav1232	Corav1232	$(CT)_8$	CCTTCACCGTTACACCCTCT	CACCAGAGAAATTCCCAACG	58	54	175	184-204
Corav1576	Corav1576	$(GA)_{11}$	CAGCCACTTCAGCTCACAAA	ACTTTCCAAGATTGCGGATT	58	54	173	186-202
Corav1859	Corav1859	$(GA)_9$	ACCTCCATGCCAGAGATGAT	CATTTTCTGGCCCCTTCTTT	56	53	188	202-219
Corav2208	Corav2208	$(GA)_{10}$	TGTGTTCCATTGGTCTCAGC	CATTGGAAGAACTCCCCTGA	56	53	184	201-237
Corav2241	Corav2241	$(CT)_8$	ACCCACGAGAAATTGGAGTG	GGGGAAGTGGGGATAGAAAG	58	54	231	238-253
Corav2387	Corav2387	$(CT)_8$	AGGGTTCCCAGTGATGACAC	TTCTTCGATCCCTCTTCTTCA	56	53	216	225-239
Corav2560	Corav2560	$(CT)_{11}$	TCCTTCTCCTCTCCCTCCTC	CACACGGTAACAAAGGCAAA	58	54	166	177-197
Corav2564	Corav2564	$(CT)_9$	CCCAGTCCCCCTTATAAACC	TGGGTGGAGATTTTGGAAAG	60	54	171	179-191
Corav4911	Corav4911	$(TA)_8$	CGCTTGTATGTCACCCTTCC	GCCTACGAAAAGATCGCTTG	56	53	177	178-204
Corav6822	Corav6822	$(TA)_9$	ATGGGTGAGAGAGACCTGGA	AGCACTTTGAAACACCACCA	58	54	164	179-189
Corav692	Corav692	$(AG)_9$	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

Marker	**N_G**	**N_A**	**H_e**	**H_o**	**PIC**
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
Mean	5.90	4.60	0.56	0.57	0.50

539

540

Table 3 - Distribution of the segregation types expected in the F_1 population 'TGL' x 'Merveille de Bollwiller' (syn 'Hall's Giant') and putative 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 M indicates markers that segregate from the female and male parents, respectively

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	*bes*1 *bzr*1 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

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

Marker	*C. americana* (n = 2)	*C. cali fornica* (n = 2)	*C. chinensis* (n = 2)	*C. colurna* (n = 2)	*C. cornuta* (n = 1)	*C. heterophylla* (n = 1)	*C. fargesii* (n = 1)	*C. ferox* (n = 1)	*C. sieboldiana* (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
Total (%)	88.9	75	80.6	11.1	5.6	69.4	83.3	63.9	11.1	

547