Discovery, characterization and linkage mapping of simple sequence repeat markers in hazelnut

Gehendra Bhattarai¹ and Shawn A. Mehlenbacher²
Department of Horticulture, Oregon State University, Corvallis, OR 97331

Number of figures: 2, plus 1 supplemental figure
Number of tables: 3, plus 6 supplemental tables

Received for publication ________________, Accepted for publication ________________.

This research was supported by the Oregon Hazelnut Commission, USDA-NIFA Specialty Crops Research Initiative Competitive Grants (2009-51181-06028 and 2016-51181-25412), and a USDA Specific Cooperative Agreement for Eastern Filbert Blight Research.

This research partially fulfilled the requirement for G. Bhattarai’s Master of Science degree. The authors thank Erik Rowley and Todd Mockler for providing the hazelnut genome sequences used in this study.

¹Former graduate research assistant. Current address: Department of Horticulture, University of Arkansas, Fayetteville, AR 72701

²Corresponding author. E-mail: shawn.mehlenbacher@oregonstate.edu
Abstract. From the genome sequence of European hazelnut (*Corylus avellana* L.), 192 new polymorphic simple sequence repeat (SSR) markers were developed, characterized, and used to investigate genetic diversity in 50 accessions. Next-generation sequencing allows inexpensive sequencing of plant genomes and transcriptomes, and efficient development of polymorphic SSR markers, also known as microsatellite markers, at low cost. A search of the genome sequence of 'Jefferson' hazelnut identified 9094 fragments with long repeat motifs of 4, 5 or 6 bp, from which polymorphic (SSR) markers were developed. The repeat regions in the ‘Jefferson’ genome were used as the references, to which genomic sequence reads of seven additional cultivars were aligned in silico. Visual inspection for variation in repeat number among the aligned reads identified 246 as polymorphic, for which primer pairs were designed. PCR amplification followed by agarose gel separation indicated polymorphism at 195 loci, for which fluorescent forward primers were used to amplify DNA of 50 hazelnut accessions. Amplicons were post-PCR multiplexed for capillary electrophoresis, allele sizes were determined for 50 accessions. After eliminating three, 192 were confirmed as polymorphic, and 169 showed only one or two alleles in each of the 50 cultivars, as expected in a diploid. At these 169 SSRs, a total of 843 alleles were found, for an average of 4.99 and a range of 2 to 17 alleles per locus. The mean observed heterozygosity, expected heterozygosity, polymorphism information content, and the frequency of null alleles were 0.51, 0.53, 0.47, and 0.03, respectively. An additional 25 primer pairs produced more than two bands in some accessions with an average of 6.8 alleles. The UPGMA dendrogram revealed a wide genetic diversity and clustered the 50 accessions according to their geographic origin. Of the new SSRs, 132 loci were placed on the linkage map. These new markers will be useful for diversity and parentage studies, cultivar fingerprinting,
marker-assisted selection (MAS), and aligning the linkage map with scaffolds of the genome sequence.

Additional index words: filbert, *Corylus avellana*, microsatellite, genetic diversity, dendrogram

European hazelnut (*Corylus avellana* L.), one of the world’s major nut crops, is an important crop in the Willamette Valley of Oregon where 99% of the U.S. hazelnut crop is produced. Over a 10-year period (2005-14), Turkey produced 67.2% of the world crop, followed by Italy (12.6%), the USA (4.0%), Azerbaijan (3.4%), and Georgia (3.1%) (Food and Agriculture Organization of the United Nations, 2017). In Oregon, hazelnuts are planted on 27,000 ha and plantings are expanding at 1,200 ha per year (Oregon Hazelnuts, 2018). There are also many new plantings and production is increasing in China, Chile, Georgia and other countries.

Hazelnut is diploid (2n = 2x = 22) and belongs to the family Betulaceae. Of the 13 commonly recognized species of *Corylus*, only the European species is widely planted for commercial production. Hazelnuts are monoecious, dichogamous, and self-incompatible, and thus cross-pollinated in nature. Cultivars are highly heterozygous and clonally propagated. Most of the world’s cultivars are selections from local wild vegetation. Based on simple sequence repeat (SSR) markers, most cultivars have been assigned to one of four major geographical groups: Central European, Black Sea, English or Spanish-Italian (Boccacci et al., 2006; Gökirmak et al., 2009).
Simple sequence repeat markers are DNA segments comprised of tandem repeat motifs that are 1-8 nucleotides in length. SSRs are polymerase chain reaction (PCR) based markers for which primers are designed from conserved sequences that flank the repeat region. At polymorphic loci, PCR amplification results in different product sizes. SSR markers are preferred over several other marker types because of their high level of polymorphism, multiple alleles, co-dominant inheritance, good genome coverage, interspecific and inter-generic transferability, high reproducibility, amenability to automated high-throughput genotyping, and ease of sharing among laboratories (Parida et al., 2009). SSR markers are widely used in plant genetics and breeding, including studies of genetic diversity and evolution, association, gene flow, genetic mapping, gene tagging, gene cloning, cultivar identification, parentage analysis, identification of duplicates, marker-assisted selection (MAS), and quantitative trait locus (QTL) analysis (Ellegren, 2004; Hearne et al., 1992; Parida et al., 2009). In hazelnut, SSR markers were initially developed from DNA libraries enriched for specific repeats (Bassil et al., 2005a, 2005b, Boccacci et al., 2005; Gürcan et al., 2010). Later, Gürcan and Mehlenbacher (2010a) developed SSRs from the sequences of inter simple sequence repeat (ISSR) markers and flanking regions. Sathuvalli and Mehlenbacher (2013) developed SSR markers from bacterial artificial chromosome (BAC) sequences. SSRs have also been developed from sequences of the hazelnut transcriptome (Colburn et al., 2017) and genome (Bhattarai and Mehlenbacher, 2017), and sequences in public databases (Boccacci et al., 2015; Gürcan and Mehlenbacher, 2010b). Although about 525 SSR markers are publicly available in hazelnut, the majority contain di-nucleotide motifs which suffer from PCR artifacts that prevent automated scoring (Testolin and Cipriani, 2010). In this study, we mined the 'Jefferson' genome sequence to develop new
polymorphic SSR markers with repeat motifs of 4, 5 or 6 bp and further saturated the reference hazelnut genetic linkage map.

**Materials and Methods**

**Plant Material.** A set of 48 diverse hazelnut accessions plus the two parents of a mapping population (Table 1) were used to characterize the new SSR markers. These accessions were a subset of the 198 unique types investigated by Gökirmak et al. (2009) and are phenotypically, genetically and geographically diverse. The same 50 accessions were used in previous characterization studies (Gürçan and Mehlenbacher, 2010a; Gürçan et al., 2010; Colburn et al., 2017; Bhattarai and Mehlenbacher, 2017). Of these accessions, 24 were used to screen all primer pairs for polymorphism on agarose gels. For SSRs that appeared to be polymorphic on agarose gels, all 50 accessions and capillary electrophoresis were used to characterize them.

**DNA Extraction.** DNA was extracted from two to four young leaves collected during early spring from trees growing in fields at the USDA-ARS National Clonal Germplasm Repository (NCGR) and Smith Horticultural Research Farm of Oregon State University (OSU) in Corvallis, OR. DNA extraction followed the method of Lunde et al. (2000) without RNAase treatment. The extracted DNA was quantified using ultra-violet spectrophotometry with a BioTek Synergy 2 Multi-Mode Reader with a Take 3 microplate reader and the data was analyzed with Gen5 software (Biotek Instruments, Winooski, VT). The DNA was diluted with TE buffer to a concentration of 20 ng·µL⁻¹.

**SSR Identification.** The hazelnut genome sequences of Rowley (2016) generated by Illumina technology were used in this study. The 'Jefferson' genome sequence (93× coverage with 333,492 sequences comprising 224.58 Mb) was used as the reference. Seven other cultivars (Barcelona, Ratoli, Tonda Gentile delle Langhe, Tonda di Giffoni, Daviana, Hall's Giant and
Tombul) were re-sequenced at lower coverage (~20×). The 'Jefferson' genome sequences were searched for SSRs using the MISA tool (Thiel et al., 2003) with the search criteria for minimum number of repeats set at six for di-nucleotide repeats, five for tri-nucleotide repeats, and four each for tetra-, penta-, hexa-, hepta- and octa-nucleotide motifs. This study pursued only the tetra-, penta-, and hexa-nucleotide repeat SSRs. Repeats that contained only As and Ts were removed, as they tend to be very difficult to score. Short fragments (< 400 bases) and fragments with repeats near the ends of the contigs were also removed, the latter because they lacked a flanking sequence sufficient for primer design. Paired-end Illumina reads from the seven other cultivars were concatenated into a single file, and aligned with SSR-containing sequences in the ‘Jefferson’ reference using the MAQ program (Li et al., 2008). The aligned reads were imported into Tablet software (Milne et al., 2013) for visualization and identification of SSRs with variation in the number of repeat units but conserved flanking sequences. Tablet software displayed the reference sequence at the top of the screen, with the aligned reads from the seven cultivars shown in subsequent rows without identification of the originating cultivar. The four nucleotides were displayed with unique colors for ease of visualization. After alignment and visualization, each sequence was classified as “clearly polymorphic”, “slightly polymorphic”, “not polymorphic”, “poor alignment” or “no reads”. Of these categories, only the “clearly polymorphic” sequences were pursued, and forward and reverse primers were designed using the Websat (Martins et al., 2009) and Primer 3 (Untergasser et al., 2012) programs with lengths of 18 to 27 bp, annealing temperature of 60 °C, and a range of expected product sizes of 90-400 bp to facilitate post-PCR multiplexing. Non-fluorescent forward and reverse primers were ordered from Integrated DNA Technologies (Coralville, IA). The "slightly polymorphic" sequences, for which <2% of the aligned reads showed variation in the number of repeats, were not pursued.
PCR Amplification for Polymorphism Screening. PCR was performed with each primer pair using DNA of 24 accessions in the diversity panel (Table 1). PCRs were in 10-µl volumes containing 0.3 µM each of forward and reverse primers, 1× Biolase NH₄ reaction buffer, 2 mM MgCl₂, 200 µM each of dATP, dCTP, dGTP, and dTTP, 20 ng template DNA, and 0.25 units of Biolase DNA polymerase (Bioline USA Inc., Taunton, MA). PCR amplification was performed in GeneAmp PCR system 9700 thermal cyclers (Applied Biosystems, Foster City, CA) in 96-well plates using the following program: denaturation at 95 °C for 5 min followed by 40 cycles of at 94 °C for 40 s, 60 °C for 40 s, 72 °C for 40 s, a final extension at 72 °C for 7 min, and an infinite hold at 4 °C. The PCR products were separated by electrophoresis on 3% agarose gels in TBE buffer at 90V for 3.5 h, stained for 30 min in ethidium bromide and then destained in distilled water for 25 min. The gels were then photographed under ultraviolet light using a BioDoc-It® Imaging System (UVP, Upland, CA), and the images were inspected for size polymorphism among the 24 genotypes.

Genotyping at Polymorphic SSRs. For SSR loci that were polymorphic on the agarose gels, fluorescent forward primers labeled with 6-FAM or 5-HEX were ordered from Integrated DNA Technologies and fluorescent forward primers labeled with NED were ordered from Applied Biosystems. The fluorescent forward and non-fluorescent reverse primers (Supplemental Table S1) were used to amplify DNA from the 48 accessions and the two parents of the mapping population (Table 1). The use of fluorescent primers and PCR products of different sizes allowed efficient post-PCR multiplexing of 5-7 primer pairs in a single well. For multiplexing, 2 µL of the PCR products from each primer pair were mixed and diluted with water to make a final volume of 150 µL. An aliquot of 1-1.5 µl of multiplexed PCR products was submitted to the Core Labs of OSU’s Center for Genome Research and Biocomputing (CGRB) for fragment
sizing by a capillary electrophoresis instrument(ABI 3730, Life Technologies, Carlsbad, CA)
using ROX-500 as the size standard. GeneMapper® software (Life Technologies, Carlsbad, CA)
was used for allele sizing and sizes were recorded in a spreadsheet. PCR amplification and
capillary electrophoresis were repeated if the initial PCR failed or the result was ambiguous, and
to verify unique allele sizes.

Characterization of Polymorphic SSRs

For polymorphic primer pairs that showed the expected one or two PCR products in all
cultivars, PowerMarker software (Liu and Muse, 2005) was used to calculate number of alleles
(n), observed heterozygosity ($H_o$), expected heterozygosity ($H_e$) and polymorphism information
content (PIC). Observed heterozygosity is the number of heterozygous genotypes divided by the
total number of genotypes. Expected heterozygosity estimates the probability that a randomly
chosen individual is heterozygous at a locus and is calculated as $H_e = 1 - \sum_{i=1}^{n} p_i^2$, where $p_i$
is the frequency of the $i^{th}$ allele and $n$ is the number of alleles (Nei, 1973). The polymorphism
information content (PIC) value is a measure of marker informativeness and is calculated as

$$PIC = 1 - (\sum_{i=1}^{n} p_i^2) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2p_j^2,$$

where $p_i$ is the frequency of the $i^{th}$ allele, $p_j$ is the frequency of the $j^{th}$ allele, and $n$ is the number of alleles (Botstein et al., 1980). The frequency of
null alleles ($r$) was calculated using Cervus software (www.fieldgenetics.com), which uses the
formula of Kalinowski and Taper (2006). Genetic similarities between pairs of accessions were
calculated as the proportion of shared alleles ($P_s$) and genetic distance was calculated as (1-$P_s$),
resulting in a frequency-based distance matrix. For SSRs with one or two bands per accession,
dendrograms were constructed using PowerMarker (Liu and Muse, 2005) and the unweighted
pair-group method using arithmetic averages (UPGMA) algorithms. To include SSRs with more
than two PCR products in some accessions, the genotype data for the 50 accessions at all new
SSR loci were converted to binary format (i.e. dominant scoring) using an Excel macro (Rinehart, 2004). The dendrograms were visualized using Mega7 software (Kumar et al., 2016).

**Segregation and Linkage Mapping.** For the newly developed SSR marker loci at which the parent genotypes of the mapping population (OSU 252.146 x OSU 414.062) (Mehlenbacher et al., 2006) predicted segregation, all 138 available seedlings were genotyped as described above. The expected allele sizes were known, and genotyping was in multiplex sets of 10-14 primer pairs. Each allele was scored as present or absent in each seedling, scores were tallied, and the expected ratio (1:1 or 1:1:1:1 or 1:2:1) was noted. Scores for the new SSR markers were added to those for previously mapped random amplified polymorphic DNA (RAPD) and SSR markers (Bhattarai and Mehlenbacher, 2017). Linkage maps were constructed using the two-way pseudo-testcross method and the BC1 function in JoinMap 4.1 (Kyazma, Wageningen, Netherland). Markers segregating in a ratio of 1:1 or 1:1:1:1 were scored as h, a or u for presence, absence or unknown status, respectively, for each allele in each seedling. Similarly, markers segregating in a ratio of 1:2:1 were scored as h if present in homozygous state, a if absent (homozygous) and u if heterozygous or unknown status. Heterozygotes at loci segregating in a 1:2:1 ratio provide no information useful for mapping. Markers were assigned to linkage groups (LGs) using a median logarithm of odds (LOD) score of 12 (range 9 to 15). Maps were constructed separately for each linkage group using the maximum likelihood (ML) algorithm and distances shown in Haldane units (cM). Markers that clustered loosely with the others and fell out at LOD scores <9 were removed. Markers present in repulsion phase were included in the map by creating "dummy variables", whose use allowed the merger of the repulsion phase and coupling phase markers, and generation of a single map for each linkage group in each parent. The JoinMap output was inspected for “Fit and Stress” and markers were removed in stepwise
manner until the "Nearest Neighbor Stress" value for all markers was less than an arbitrarily set value of 7.60. The "Nearest Neighbor Fit (cM)" values were also inspected, as high values indicate blocks of markers that fit poorly with adjacent markers. In addition, the length of the gaps between markers was inspected, with gaps of >20 cM considered suspicious.

Results

SSR Discovery and Characterization. A search of the 'Jefferson' genome sequence identified 51,268 SSRs with repeat motifs of two to eight bp (Table 2). Di-nucleotide repeats (64.90%) were most abundant, followed by tri-nucleotide (16.99%), tetra-nucleotide (12.49%), penta-nucleotide (3.43%), hexa-nucleotide (1.82%), hepta-nucleotide (0.37%), and octa-nucleotide (0.001%) repeats. The number of repeats in these SSRs ranged from 6 to 14 in the di-nucleotide repeats, and decreased as the motif length increased. Among the di-nucleotide repeats, AG and AT repeats were most common, representing 49.4% and 40.2% of the total, respectively (Supplemental Table S2). Among the tri-nucleotide repeats, AAT and AAG were most common, representing 43.1% and 28.4% of the total, respectively. The most common motifs were AAAT (55.0%) among the tetra-nucleotide repeats, AAT (36.3%) and AAAAG (25.7%) among the penta-nucleotide repeats, and AAAAAG (18.3%) and AAAAAT (12.6%) among the hexa-nucleotide repeats. The focus of this study was the tetra-, penta- and hexa-nucleotide repeats, for which the maximum number of repeats was 8, 6 and 6, respectively. Among the 192 polymorphic SSRs markers developed in this study, the most common motifs were ACGT/TGCA, AAAGG/TTTCC and AGTTGG/TCAACC for the tetra-, penta- and hexa-nucleotide repeats, respectively.
Initially, the number of sequences containing tetra-, penta-, and hexa-nucleotide repeats was 9094. The number of sequences was reduced in stepwise manner, resulting in the final set of 192 polymorphic SSR markers (Table 3). Of the 9094, 4774 were removed because the repeat motifs contained only As and Ts. The removal of short fragments (< 400 bp) and those with the repeats at the fragment ends further reduced the number to 1013. After alignment with the reference genome, 246 were scored as "polymorphic" as they showed variation in number of repeat units but conserved flanking sequences. An additional 17 were scored as "slightly polymorphic", and 616 showed no polymorphism in number of repeat motifs. For 114 sequences, the reads of the seven cultivars aligned poorly with the reference, and for 20 sequences none of the reads from the seven cultivars aligned with the reference. Primers were designed for the 246 scored as polymorphic based on visual inspection of aligned sequences. After amplification of 24 diverse accessions followed by separation on agarose gels, 195 were scored as polymorphic. Fluorescent forward primers were ordered for these 195, which were then used to amplify and score allele sizes in 50 accessions after separation by capillary electrophoresis. Two primer pairs gave poor amplification in many cultivars, and one was monomorphic, while 192 new polymorphic SSR markers were pursued, of which 89 were labeled with 6-FAM, 70 with 5-HEX, and 33 with NED (Supplemental Table S1). Of the newly developed SSRs, 60, 66 and 66 consist of tetra-, penta-, and hexa-nucleotide repeats, respectively. Allele sizes at the 192 new marker loci in the 50 accessions are presented (Supplemental Tables S3 and S4). At 169 SSR loci (88.0%), all 50 accessions showed the one or two alleles expected in a diploid, and diversity parameters were estimated for them (Table 3 and Supplemental Table S3). However, at 23 SSRs (12.0%), more than one accession generated more than two PCR products; for these the presence of each allele was scored in each accession
Hazelnut simple sequence repeat markers

In the 169 characterized markers (Table 3), 846 alleles were detected. The number of alleles per locus ranged from 2 to 17 with an average of 4.99. Expected heterozygosity had a mean of 0.53 and a range of 0.04 to 0.85, while observed heterozygosity had a mean of 0.51 and a range of 0.04 to 0.90. The PIC value averaged 0.47 and ranged from 0.04 to 0.84. GB726 was the most polymorphic marker ($H_o = 0.81$), while the four least informative markers (GB404, GB405, GB446 and GB616) had PIC values less than 0.10. Of these, GB405 had four alleles while the other three had only two alleles each. The average frequency of null alleles was 0.03, with values for individual loci ranging from -0.20 to 0.59. At the 23 primer pairs that gave more than two PCR products in one or more accessions, 170 alleles were present with an average of 7.39 alleles per primer pair (Table 3). Allele size distributions at the 192 polymorphic SSR markers are shown as histograms (Supplemental Fig. S1).

**Segregation and Linkage Mapping.** A total of 135 loci segregated in the mapping population, of which 70 (51.9%) segregated in a 1:1 ratio, 35 (25.9%) in a 1:1:1:1 ratio, and 30 (22.2%) in a 1:2:1 ratio (Supplemental Table S5). New markers were placed in each of the 11 linkage groups, and the resulting genetic map included 132 new and 297 previously-mapped SSR loci (Fig. 2). Three loci (GB733, GB737 and GB746) showed severe segregation distortion and could not be assigned to a LG (Supplemental Table S4). The lengths of the linkage groups ranged from 63.5 to 277 cM.

**Genetic Diversity in Hazelnut Accessions,** A dendrogram was constructed for the 50 accessions using the UPGMA algorithm (Fig. 1) and allele sizes at 192 loci. The dendrogram confirmed the wide genetic diversity in hazelnut, grouped the accessions according to their geographic origins, and showed the four major groups seen in previous studies: Spanish-Italian, Black Sea, Central European, and English (Bhattarai and Mehlenbacher, 2017; Boccacci et al.,
2006; Colburn et al., 2017; Gökirmak et al., 2009; Gürcan et al., 2010). The Central European group includes four cultivars of German origin ('Gunslebert', 'Hall's Giant', 'Gustav's Zellernuss' and 'Early Long Zellernuss' as well as 'Bergeri' from Belgium and 'Alli' from Estonia. The Black Sea group includes six accessions from Turkey and one from Georgia. The English group includes 'Cosford' from Reading, England, 'Rote Zellernuss' from the Netherlands, and 'Buttner's Zellernuss' from Landsberg, Germany. Their adjacent placement in the dendrogram suggests that the latter two might be seedlings of 'Cosford'. The ornamental 'Contorta' originated in England, as did 'Du Chilly', which is grown as 'Kentish Cob' in southeastern England. 'Pellicule Rouge', obtained from France, has leaves and pellicles with anthocyanin, but all other traits and its DNA fingerprint are identical to the English cultivar 'White Filbert' (Gökirmak et al., 2009).

'Des Anglais' was obtained from Bordeaux, France, but its name indicates an English origin. All of the Spanish and Italian cultivars were included in the large Spanish-Italian group.

Discussion

The high throughput and low cost of next-generation sequencing allows plant genomes to be sequenced and SSR markers to be developed efficiently as a by-product of genome sequencing. In hazelnut, we identified 9094 'Jefferson' genome sequences that contained tetra-, penta- or hexa-nucleotide repeats and proceeded in a stepwise manner to develop polymorphic SSR markers from them. Of the 1,013 sequences for which aligned sequences were visually inspected, 246 showed variation in number of repeat units but conserved flanking sequences. After primer design, amplification, allele sizing and characterization, this number was reduced to the 192 new polymorphic SSR loci described in this paper, of which 169 amplified only one or two bands in all cultivars, as expected in a diploid species. The ratio of single-locus polymorphic SSRs developed to number of primer pairs designed was 68.7% in this study,
compared to 28.5% for the mostly tri-nucleotide repeat SSRs developed from transcriptome sequences (Colburn et al., 2017) without in silico evaluation. The comparison illustrates the improved efficiency that results from this evaluation. For the di-nucleotide repeat SSRs developed from enriched libraries, the ratio was an intermediate 57% (Gürcan et al., 2010).

Although single nucleotide polymorphism (SNP) markers can be generated in large numbers by procedures including genotyping-by-sequencing (GBS) (Elshire et al., 2011) and double digest RADseq (Peterson et al., 2012), SSR markers continue to have many research and breeding applications in plants. In hazelnut, for example, Özturk et al. (2017) studied associations of SSR markers with nut and kernel traits in a Slovenian hazelnut collection, and Marinoni et al. (2018) used SSR and SNP markers to identify quantitative trait loci (QTL) in a seedling population from 'Tonda Gentile delle Langhe' x 'Merveille de Bollwiller'.

Most of the tetra-, penta-, and hexa- SSRs identified in this study were composed of only As and Ts (52%). The very high frequency of A and T bases in SSRs have been observed in many other tree species (Xia et al., 2017). Mono-nucleotide repeats, although abundant in the hazelnut genome, were not pursued as they tend to be difficult to score. Di-nucleotide repeats were not pursued as large numbers had previously been developed (Bassil et al., 2005a, 2005b; Boccacci et al., 2005, 2015; Gürcan et al., 2010, Gürcan and Mehlenbacher, 2010a; Sathuvalli and Mehlenbacher, 2013) and stuttering is a common problem with di-nucleotide repeats (Hearne et al., 1992). Colburn et al. (2017) developed 111 SSRs from transcriptome sequences, most of which were tri-nucleotide repeats, and Bhattacharai and Mehlenbacher (2017) developed 150 tri-repeat SSRs from the Illumina draft genome sequence for 'Jefferson'. This study adds 192 markers to the 525 previously developed, and further saturates the genetic map. Previous studies (Bassil et al. 2005a, 2013; Boccacci et al., 2005, 2015; Gürcan and Mehlenbacher,
2010b) showed a high rate of transferability to other Corylus species, and lower transferability (~30%) to the genera Betula and Alnus, which are also in the family Betulaceae. Thus, the SSR markers developed in this study could have applications beyond C. avellana.

Of the new polymorphic SSRs developed in this study, 23 primer pairs amplified more than two bands in some hazelnut accessions. SSRs of this type were previously reported in hazelnut (Bhattarai and Mehlenbacher, 2017; Gürcan and Mehlenbacher, 2010a) and other plant species (Chavarriaga-Aguirre et al., 1998; Kim et al., 2014). We attribute this to either non-specific amplification or genome duplication events, including paralogous sequences.

The count of private or unique alleles (that occur in only one individual) is 201 at 103 SSR loci (Supplemental Table S6). These unique alleles were confirmed by repeating the PCR and allele sizing. Null alleles, which fail to amplify with PCR, are the result of a mutation in either one or both primer binding sites of the genome sequences. The presence of null alleles leads to errors in pedigree and segregation analysis (Dakin, 2004; Pemberton et al., 1995). The mean frequency of null alleles was low (r = 0.03), although a few loci had high values and 16 (GB405, GB409, GB421, GB454, GB461, GB465, GB505, GB532, GB602, GB624, GB640, GB646, GB649, GB671, GB718 and GB721) had values > 0.20.

The UPGMA dendrogram based on binary data of allele sizes at 192 SSRs showed accessions clustered according to their geographic origin into four main groups: Central European, Black Sea, English, and Spanish-Italian. This grouping is similar to previous studies (Boccacci et al., 2006; Colburn et al., 2017; Gökirmak et al., 2009; Gürcan et al., 2010). The dendrogram (Fig. 1) shows a high level of diversity in the 50 accessions, including 13 diverse accessions at the top that did not cluster with the any others (Fig 1).
Of the 192 new SSR markers, 135 segregated in the mapping population, of which 132 were placed on the map and three with distorted segregation remained unlinked. Segregation ratios at 135 loci (Supplemental Table S4) were examined for goodness-of-fit to expected ratios. Poor fit to Mendelian expectations \( (P < 0.05) \) was observed at 20 loci, of which 11 showed very poor fit \( (P < 0.01) \). Several SSRs did not segregate in the reference mapping population. Their map locations may be determined in future research using alternate mapping populations, or by alignment with a mapped genome sequence contig. Linkage groups are presented separately for the female and male parents (Fig. 2). The female parent map consists of 11 distinct linkage groups, while the male parent map initially had 10 distinct groups, with LG2 and LG7 merged into a single large group, as first encountered by Mehlenbacher et al. (2006) and more recently by Beltramo et al. (2016) for a population from 'Tonda Gentile delle Langhe' x 'Merveille de Bollwiller'. Reciprocal translocations are common in hazelnut (Salesses, 1973; Salesses and Bonnet, 1988), and may be responsible for this merger. We separated LG2 and LG7 in the male parent map following the method of Colburn et al. (2017) and present 11 groups. The newly developed markers were assigned to all 11 LGs (Fig. 2), but were not evenly distributed.

Four markers (GB478, GB456, GB605 and GB671) were placed on LG6 <20 cM from the locus from 'Gasaway' for resistance to eastern filbert blight (EFB) caused by *Anisogramma anomala* (Peck) E. Müller (Mehlenbacher et al., 2006), although none are closer than four SSRs developed from bacterial artificial chromosome (BAC) sequences that co-segregate with the allele for resistance (Sathuvalli and Mehlenbacher, 2013). EFB resistance from five other sources maps to the same region (Bhattarai et al., 2017b; Colburn et al., 2017; Sathuvalli et al., 2012). Three other markers (GB467, GB482 and GB642), were assigned to LG7 in the region of the EFB resistance loci from 'Ratoli', 'Yoder #5' and *C. americana* Marsh. 'Rush' (Bhattarai et al.,
Four markers (GB609, GB618, GB669 and GB740) were placed on LG2 near the EFB resistance locus from Georgian selection OSU 759.010 (Sathuvalli et al., 2011). These new markers will be useful for mapping additional EFB resistance sources and developing breeder-friendly DNA markers in these regions for use in marker-assisted selection. Additionally, GB646 was placed on LG5 near the S-locus that controls pollen-stigma incompatibility (Ives et al., 2014; Mehlenbacher et al., 2006).

The genomic SSRs developed in this study add to those developed previously and will facilitate research in hazelnut genetics, including cultivar fingerprinting, parentage analysis, studies of genetic diversity, mapping of qualitative and quantitative trait loci (QTL), marker-assisted breeding, and the fine mapping and cloning of genes. Because SSR markers map to the same location in different populations, they allow the alignment or merger of genetic maps from different crosses. In addition, the mapped SSR markers will allow alignment of the linkage map with the genome sequences and the physical map of BAC contigs.

Tables.

Table 1. Hazelnut accessions used to characterize new simple sequence repeat markers mined from genome sequences. The first 24 were also used to screen for polymorphism on agarose gels, and all 50 were used for characterization.

Table 2. Counts of simple sequence repeats in the 'Jefferson' hazelnut genome by type and number of repeats.
Table 3. Characteristics of new simple sequence repeat markers from the genome sequence of *Corylus avellana* 'Jefferson' and linkage group assignments in the reference mapping population (OSU 252.146 x OSU 414.062).

List of Figures.

Fig. 1. UPGMA dendrogram of 50 hazelnut accessions constructed from binary data for 192 simple sequence repeat markers.

Fig. 2. Linkage maps of the hazelnut reference mapping population (OSU 252.146 x OSU 414.062). The new simple sequence repeat markers are indicated by * and bold font (PDF).

Supplemental Tables.

Supplemental Table S1. Repeat motifs, amplicon size, primer sequences and annealing temperatures of simple sequence repeat markers developed from the 'Jefferson' hazelnut genome sequence.

Supplemental Table S2. Repeat motifs at simple sequence repeats in the 'Jefferson' hazelnut genome by type and number of repeats.

Supplemental Table S3. Alleles at 169 simple sequence repeat marker loci in 50 hazelnut accessions. All cultivars showed one or two alleles at each locus.

Supplemental Table S4. Alleles at 23 simple sequence repeat marker loci in 50 hazelnut accessions. Cultivars showed more than two alleles at more than one locus.

Supplemental Table S5. Segregation at simple sequence repeat marker loci in the hazelnut reference mapping population (OSU 252.146 x OSU 414.062).

Supplemental Table S6. Private (unique) allele sizes at simple sequence repeat marker loci.
List of Supplemental Figure

Supplemental Fig. S1. Histograms showing allele frequencies at 194 new simple sequence repeat loci in hazelnut (PDF).

Literature Cited


doi.org/10.1371/journal.pone.0195408

genetic linkage map for hazelnut (Corylus avellana L.) based on RAPD and SSR markers. Genome
49:122-133. doi:10.1139/g05-091

Using Tablet for visual exploration of second-generation sequencing data. Briefings Bioinformatics

70:3321-3323.


Molecular genetic diversity and association mapping of nut and kernel traits in Slovenian hazelnut


Kumar, S., G. Stecher, and K. Tamura. 2016. MEGA7: Molecular Evolutionary Genetics Analysis

Li, H., J. Ruan, and R. Durbin. 2008. Mapping short DNA sequencing reads and calling variants using
mapping quality scores. Genome Res. 18:1851–1858. doi:10.1101/gr.078212.108


Marinoni, D.T., N. Valentini, E. Portis, A. Acquadro, C. Beltramo, S.A. Mehlenbacher, T.C. Mockler,
E.R. Rowley, and R. Botta. 2018. High density SNP mapping and QTL analysis for time of leaf

doi.org/10.1371/journal.pone.0195408


genetic linkage map for hazelnut (Corylus avellana L.) based on RAPD and SSR markers. Genome
49:122-133. doi:10.1139/g05-091

Using Tablet for visual exploration of second-generation sequencing data. Briefings Bioinformatics

70:3321-3323.


Molecular genetic diversity and association mapping of nut and kernel traits in Slovenian hazelnut

Hazelnut simple sequence repeat markers


