Conservation Genetics Resources
https://doi.org/10.1007/s12686-020-01157-5

TECHNICAL NOTE

Microhaplotype genotyping-by-sequencing of 98 highly polymorphic markers in three chestnut tree species

Benoit Laurent[1] · Clément Larue[1,2] · Emilie Chancerel[1] · Erwan Guichoux[1] · Rémy J. Petit[1] · Teresa Barreneche[3] · Cécile Robin[1] · Olivier Lepais[1]

Received: 27 January 2020 / Accepted: 4 June 2020

Abstract
Chestnut species have large ecological, cultural and economic importance. Developing genetic markers for these species is of interest for conservation, breeding or evolutionary studies. We designed 192 primer pairs targeting microsatellites detected in the *Castanea mollissima* reference genome and tested them on *C. sativa* and *C. crenata*. We PCR amplified 3 × 50 microsatellites in 106 chestnut trees. Microhaplotype calling accounting for all polymorphisms resulted in a total of 98 high confidence polymorphic markers. Mean number of haplotypes per marker was 9.05 with respectively 71%, 12% and 16% of the variation corresponding to microsatellite variation in repeats number, SNP within the repeat motif and SNP or INDEL in the flanking sequence. Overall, the simple protocol described here generated a powerful multilocus genetic dataset for chestnut genetic investigations.

Keywords *Castanea sativa* · *C. crenata* · *C. mollissima* · Microsatellites · SNP · INDEL · SSR-seq

Introduction

The genus *Castanea* Mill. (chestnuts) comprises at least eight interfertile species (The Plant List 2013) including *Castanea dentata* (Marshall) Borkh from North America, *C. sativa* Mill. from Europe, and *C. crenata* Siebold & Zucc and *C. mollissima* Blume from Asia (Pereira-Lorenzo et al. 2012). Very closely related to oaks (Kremer et al 2012), another *Fagaceae* genus, chestnuts are keystone multipurpose trees that provide habitat for wildlife, edible nuts for both wildlife and human, timber and tannins, and play important cultural roles in some human societies (Pereira-Lorenzo et al. 2012; Powell et al. 2019). But chestnuts have endured one of the greatest environmental disasters that led to the functional extinction of *C. dentata* in North America and to important reduction of *C. sativa* population in Europe as a consequence of invasive pathogens (Desprez-Loustau et al. 2007; Powell et al. 2019). In contrast, Asian chestnuts harbor quantitative resistance to these pathogens, allowing resistance breeding (Barreneche et al. 2019). Improved knowledge on the genetic diversity of chestnuts requires the development of numerous, reliable and polymorphic molecular markers. SSRs have remained popular markers given their high polymorphism, reproducibility, transferability and ease of detection (Guichoux et al. 2011; Lepais and Bacles 2011). Hence, the recent development of sequence-based microsatellite genotyping approaches (e.g. Vartia et al. 2016) has raised much interest. The aim of this study was to setup a sequence-based genotyping method for *C. sativa*, *C. crenata* and *C. mollissima*, three chestnut species of major economic and ecological importance.

SSR marker design and genotyping were conducted using the workflow described in Lepais et al. (2020). A total of 196 primer pairs targeting SSR markers were identified from the *C. mollissima* reference genome v4.1 (Staton et al. 2019) using QDD v3.1 (Meglécz et al. 2014). Briefly, out of the 125,824 microsatellites identified across the genome, 39,473 had successful primers designed with a targeted predicted amplicon between 120 and 200 bp and stringent parameters to increase multiplexing success (Lepais et al. 2020). Only primer pairs with no homopolymer, targeting no other microsatellite, with no nanosatellite in primers or in the flanking

✉ Benoit Laurent
benoit.laurent@inrae.fr

[1] Univ. Bordeaux, INRAE, BIOGECO, 33610 Cestas, France

[2] INVENIO, Maison Jeannette, 24140 Douville, France

[3] Univ. Bordeaux, INRAE, BFP, 33140 Villenave d'Ornon, France

Published online: 09 June 2020

Table 1 SSR-markers genotyped on the 106 trees

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS001	CCTGGG AAAGGT GACAGT TTCAGGT	ACCTTTGCT GTTCCC ACCTTT GCCG	P 2	99	268,104	FL 1	7	7	0	0	6%	0%
CS005	GGTGGGCTA GATATT CCTTCC TCTGGT	GCTAAGGAT GAGCAT ACATCG CTAA	P 2	156	183,848	FL 2	8	6	3	3	6%	0%
CS007	TCCCAAAGC GATCTA CCATGG AAGCT	TGCACCACA TAACTT CCCTAG GTACCC	P 1	348	32,105	RF 2	9	5	2	–	6%	0%
CS008	TGTTCACTT CCAATG AGAACC CATCCT	GCACGTATC GAATTA AACACT TAGAGG	P 3	421	91,665	RF 2	5	5	0	–	13%	0%
CS009	TTCGAGAAG CTTTCT GCCTTG CTCA	CGAAAC AAACGG AGCGTA AATTTC CCA	P 3	444	192,690	FL 2	12	11	0	1	4%	3.8%
CS010	CGACAATTT GCAATC TTCTCC ACCCA	GGCTGCCTA AACAAG CTAAGT TATGAGCA	P 3	500	214,240	RF 2	4	4	0	–	17%	0%
CS012	ACTGTGGGC GTGGAT GGACTT GACT	ACCAAA CAAACG GGCCTC ACAGTGC	P 3	592	24,216	RF 2	10	9	1	–	5%	0%
CS016	AGGAGTATA ACTTTG TAGGCT GTCCT	TGGGCA TGAAGG AAGGGA GGAAGGA	P 3	1043	20,533	FL 2	4	3	0	1	3%	0%
CS017	GCAGACTGA TGCGAA TTCTTT AACT	ATTGCAGAC AGCTCC ACCAGG TCCT	P 1	1105	23,737	RF 1	22	13	6	–	6%	3.8%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS019	ACACATTTACAACCACTGCCTATGCT	GGGCTTGCCTTTCCTCATCAGATGCA	P 1	1368	55,781	RF 1	12	12	0	–	7%	0%
CS022	AATGAATGTGAGCCACGCGACGGTT	GTGCCTGGCATGCCCTGTACATCTT	P 1	2054	8759	FL 2	10	8	2	1	7%	0%
CS025	AGGGCGTTCCAATCACAATGAAACA	ACCAGGCAAGGTGGTATGCACTCCA	P 3	2519	33,044	FL 2	26	13	0	11	21%	3.1%
CS026	GTCAAGACAATGCATCTCAATAAAGCC	GCGTCTGCGTGATAGATTAGAAGTT	P 3	2718	74,398	RF 2	11	9	2	–	18%	0%
CS027	AACCAATTCGCCTGGGCCTGTG	TCCTCCGCCACACGTTCCGTACTAC	P 3	2949	13,306	RF 2	5	4	1	–	11%	0%
CS028	CGGAACTCAAGATGGGCAAGGGAGC	CGGAAGTGAGACAAGGTGCATGATCGG	P 2	3267	31,354	FL 2	13	6	2	5	6%	0%
CS029	CGATGTCGGCCTGTTCACCCACCTA	GTGAACTGACCGTGCGTTCTGGGAC	P 3	3577	59,792	RF 1	12	8	4	–	6%	0%
CS031	ACCTACCGCTCCACAAACCTTGGCA	ATCTGTCGTATTGGCATCTGAAGCA	P 2	5447	636	RF 2	13	8	3	–	17%	0%
CS033	ACTCAAGCCTCATGAGAAATTTGTGGC	ACACCACCAAATCAAAGCAACAAGT	P 2	7957	8067	RF 2	5	5	0	–	9%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS034	GCCGCG GAGACA CCACAA ACAAGAA	ACAATCCCA CAGACA CAGTAA TAGTAG	P 1	12,748	18,087	FL 1	16	11	2	1	8%	0%
CS035	TTCCCAGTT TGTCTG CAGCAC CGTG	CCTGATGGT GAGGTT GAGCGA CTGGT	P 1	12,805	40,168	RF 2	6	5	1	–	7%	0%
CS036	GGGTGTGCA TGAATT GAATTG GATT	AGCAGTTTC CTGGAT TTCCAT TTGA	P 2	12,819	88,861	RF 2	8	7	2	–	17%	3.1%
CS037	GTCTGA TGACTC GGTCAC AGAAA	AACCCA ACCAAC CCGCCA ATACTGC	P 3	13,182	4748	FL 2	16	10	1	1	2%	0%
CS038	CTTGGACCG GTGGTT TGCTCA AGCG	TAACGG CAACTA ACTAAC GCAACGT	P 3	13,367	17,773	FL 2	7	3	3	3	0%	0%
CS039	TGCCAACCA TGACTT ATCTTG TTGAGG	TGGTTCGAG GAGGTG CGAGTA GAGC	P 2	1341	116,351	FL 2	13	6	1	10	6%	0%
CS043	TTCCACAGA GACGAA CGTGCC GAGA	CGCGGT GAAGCT GACTGT GCAGAAT	P 1	368	52,096	RF 1	9	5	1	–	7%	0%
CS047	AAGCTTATG ACATCG CGGCCC AACG	ACATGTCCA CAATCT CAGCCT TGGA	P 2	8600	345	FL 2	5	4	0	2	9%	0%
CS050	ACCCGA CAAGTC CCTAAC ATCGTCT	GGCAGT CAATTT GGCCAA GCAAACA	P 1	2548	8943	FL 2	11	5	2	3	7%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS052	ACCTGCTCT GCCCTT GTAGAA ACGC	GAGTCA CCGGAG AAGTGG GAAGCGA	P 3	3144	25,266	RF 2	8	7	1	–	8%	0%
CS053	TTAACGGTA GTGGTA ACGGCG GCGA	CTGCTCCCG CCATTT CCAACT CGTT	P 3	160	291,121	RF 2	7	6	1	–	1%	0%
CS056	TCCCACACT TTCCGA GAAACC AAAC	CGAGCTTCT AGTAGC CGCCGG ATGT	P 1	3383	2892	FL 2	13	9	3	1	7%	0%
CS057	GCTCACATG AAAGGA GGTTCA CAGCCA	GGGTGC ACTTGC CCTCTT CCTTCCT	P 2	2780	8188	FL 2	10	8	1	2	6%	0%
CS058	CACCATCGA TGCGCC GAATGT GTCC	TGGCCG AGTGTA GTAAGT CGATGGG	P 2	783	78,916	FL 2	8	5	1	2	6%	0%
CS062	TGGAGACAT ACTGAG CAATGT TCTT	TCTCTGCAA CAACAG ACTGGA CACA	P 3	12,786	64,008	RF 2	5	4	1	–	3%	0%
CS063	CCAGAGCTC ACCGAT CTCTTC GCGA	TTGAAGCAC ATTAGG GACAGC GCGG	P 2	1796	73,677	RF 1	4	4	0	–	7%	0%
CS065	TGTGAGGCA TGTTCA GGAAGT TGCA	ACAATCTGG GCAACG TTGGAA TCCA	P 3	769	33,062	FL 2	10	8	2	0	6%	0%
CS066	TCCCTGGGA CATAGT TGTGGG ATCA	TGTTGACGT GGTGCA GTTCTT GTTTG	P 3	151	198,466	RF 2	7	6	1	–	10%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS070	CCTCTGTGC GCTGTG GAGGAA TTGC	GGCAAG GGCTTG GTACTC AGCCTCT	P 2	4554	16,135	FL 2	19	7	3	9	12%	5.0%
CS071	ACGGTTGTC GATTTC AGTGTC CAGC	GGAGACCTT GGAGAT GGCGTT TCCG	P 2	2436	70,400	RF 2	18	14	3	–	9%	5.0%
CS073	ATGAAGCTG CCTTGG CTGTTG CACA	AAATTGTGG AGCAAC TGAGAT AGCT	P 3	412	122,964	FL 2	6	6	0	1	5%	0%
CS078	CTGCCATTT GGTTGG AATTCT GCAA	TGTGCAGCA GTATCG GCAATA TTGA	P 2	762	149,153	RF 2	3	3	0	–	6%	0%
CS080	GCTGAGAAT ATTGGT TCACTT TGCAGT	TGCACAGTT AACCAC ATTCAT AGCCA	P 2	1963	72,817	RF 2	6	5	1	–	10%	0%
CS081	GGTCGATCA ATCAGA CCAAAT CTCTGT	GTGACTTCT TGATGA CTATTG CTAGCA	P 3	14,268	779	FL 2	11	8	1	2	1%	0%
CS084	AGGAGT ACAAGG ACTCAC ATGCCGA	ACGGTCTAC CATTTG TAGCTT ACAC	P 1	291	108,878	RF 2	9	8	2	–	15%	2.6%
CS085	TGATGGCAT CAAAGG GATATT GCAA	TTTCTGCTT GAAATG CCTTCA TGGA	P 3	412	72,379	FL 2	6	5	2	0	3%	0%
CS086	CCCTCCCAA TCTGAG ATAACA AAGCCC	ACCCTTGCT TTACAT CTTGCT TCAA	P 1	2053	61,368	FL 1	9	6	0	5	9%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS088	TTAGGAGAT TCTGGA AATGCT GCCA	TCTTTGGTC ACATTG AAGTGC ACCC	P 1	2701	82,817	RF 2	5	4	1	–	10%	0%
CS089	GCACCAACT CCTAAA CGGCCA TTGC	TCCTGCCTT ATACAT GTCCAC TTCA	P 2	5415	38,753	FL 2	12	8	1	13	10%	2.3%
CS092	ACCCTGTGC AATAAA TCTATG CTAGCA	GCAATTTGA TGAATT TCGTGG CAAGCA	P 2	24	77,168	FL 2	7	2	0	5	6%	0%
CS099	ACAGGA AGGTTG TTCAGA CAAGACT	TTGGGA ACATGA TCAAGC GTGACCA	P 2	13,227	16,803	RF 2	9	9	0	–	6%	0%
CS103	TGCCATCAT TAGAAT GTGATC GGGT	CTCTTGATT CCTAAA CACGTA CACCT	P 2	12,735	82,249	RF 2	9	9	0	–	13%	0%
CS105	TCAAGCCAT CCATAA CTCTTT AGCCA	TGTTGGAGC GATTTC AAACAT GTGCA	P 2	1757	27,953	RF 2	6	5	1	–	6%	0%
CS106	GCAGCCTTT GCGTCA GTATTT ATGGG	TCTTCAGCT GCAATC ACCATA CACT	P 1	557	184,045	RF 1	7	7	0	–	7%	0%
CS108	GGTGTTGTC CTTCTC CTGTCG TGTT	GTGTTGGCT TTCTTT CAGTCA CTGGG	P 2	56	172,524	RF 2	2	2	0	–	6%	0%
CS111	TCTCCTTGC ACCTCC TCAACC ACCC	CGGCCT CGCGTT GAAGTA CTCGAAC	P 3	5991	15,310	RF 2	5	3	2	–	3%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS114	ACTGGA AGCAGT CAAGAA ATCCTCT	TGCAAATCA GCACTT GGGTCC TCTCA	P 2	91	5489	FL 2	13	9	0	3	10%	2.4%
CS117	TGAACA AGAAAT GTGCGG CGCAACA	CCATGAGTG CACCAC CTCTCG CTCT	P 3	173	201,681	RF 2	9	7	1	–	5%	0%
CS120	GCCTTTGCA GAGACA TCCCAT GCCG	TGGAGT GGTGGA TTGTTC ATCATG TGGG	P 2	440	140,289	RF 2	21	13	1	–	13%	2.6%
CS121	TGACACAAA GCCACG CGCATA TGCA	GCTGAACTC GCCGTA CCACCA TGGA	P 2	1189	130,827	RF 2	5	5	1	–	10%	0%
CS122	CATGAGGAA TTGGTT CGGATT GGG	TCCATCTCT GTCTGT TTCTCT GAGTTGT	P 2	2286	60,229	RF 1	9	9	0	–	23%	2.8%
CS123	ACCAAGTCA TAGCCA ACACAG CCAC	AGCATCAAC CACAAC ACCAGC AGCT	P 2	3543	47,407	FL 2	4	2	3	0	7%	0%
CS124	CATTGTAAC GCAAGC ACAGAC ATGC	GTGGCA CGAAGT CCAAAC AACGGGT	P 3	129	208,543	RF 2	15	12	2	–	11%	0%
CS126	TGGCAC CGGCTA ATTAAC CCATGTGC	TGGAGA TGCATA CAACAA TCACGGG	P 3	59	76,067	RF 2	5	5	0	–	5%	0%
CS130	AGGCCTTCA AACAAT GACCAA TGAA	AGTGAG TGCAGC TAATTG CCACTAGA	P 3	1173	141,068	RF 2	10	10	1	–	2%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS131	AGCCTTGTG CCTGTG TCCCAT GTCT	TCTGGCATA CTAATT GCTGGG ACAA	P 2	1064	123,932	FL 1	15	10	1	6	6%	0%
CS132	TGCCTGTGC ATGACT CCCGTG ATTC	AGCGCTACA TAGGTT TAAGCT AAGCG	P 2	1442	33,550	RF 2	9	9	0	–	7%	0%
CS134	AAGTGATGT GACGTC CTTCCA AGCA	TGTGCCTCA TTTCTA ATTGTT GGGT	P 2	1144	121,088	FL 2	23	16	2	2	13%	2.4%
CS136	ACAATGCTG ATCGTG TTCCTT GCTGA	AGCCTTACA CCTCAC TGGATG CTGT	P 1	4935	24,220	FL 2	11	6	2	6	7%	0%
CS137	CAGCGACAC GACCGA GCGTTA AGAC	GGGTCT CGCCGT AGATAT GGGTGCA	P 2	3999	2126	FL 2	12	7	3	14	13%	0%
CS138	ACACCCGAT TACTGC TCCTCC ACCG	TCCAACACG GTCTGA ACCTGT CGGA	P 1	207	98,933	RF 1	4	4	0	–	11%	0%
CS139	GCTGTTTCG TTAGGT GCATTG GAAGGGT	TCCTGTATC CTTTCT TGGCAA TTACTG	P 1	5513	20,994	FL 2	4	4	0	0	7%	0%
CS142	TCCAAGATG GACCTG TACGCC GTCG	GCTCCG CCGCAT TGACGT CTATTGT	P 2	30	207,466	RF 2	7	5	1	–	6%	0%
CS143	GTTGCAGCA ACCCAA GATTCA AACT	TCCGCTTAC ACTAGA TGACTC AGGC	P 1	4599	27,523	RF 2	4	3	1	–	14%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS145	TGGAGG GTGAAT GGAACT GGTGGGT	TGACCCATC TCAAAT TCAACC ATCCT	P 2	933	14,941	RF 2	9	6	2	–	14%	2.5%
CS146	TGCAATTCA TTGGTT GGGTCA AGGCA	AATCATTGG CATTAA TGAGGC TGGT	P 1	3764	29,978	RF 2	5	5	0	–	7%	0%
CS147	GGTATCTCA CTTCTT TCCTTC CACGGA	TGCTCTTGC TCACTT TCAGGA GAGT	P 3	243	202,705	RF 2	10	10	0	–	8%	2.8%
CS149	TGACCAAGT GATGCA TTAGCC TGGC	TGCTCAATC ACAAAG TCACAT GTCA	P 1	12,864	61,489	RF 2	19	11	2	–	19%	0%
CS150	CTTGTGAAC TCAGAA GCCATG GACA	AGTTGTGCA TAGCCA AGTCGA CAAA	P 1	2264	3388	FL 2	6	4	1	1	7%	0%
CS152	ACTCGGGAT CATACT TCCAGC CAAA	GGATCCTGT CCTGCA ACTTTC CTTT	P 2	223	166,331	RF 2	8	8	0	–	15%	0%
CS153	CGAGAC GGCATT GAAGCA CCCAAGT	TCGATGCGA TTTCTA CGAAGA ACCGT	P 3	1878	65,672	RF 2	4	4	0	–	20%	0%
CS157	GCCTTGATT GGTCCA AAGGTC TACA	TGGAAG AAAGAT GTTGTG AGTCTC TAG	P 2	454	122,994	RF 2	11	8	3	–	10%	0%
CS158	GGCACGCCT AAACCT AATCCG ACCC	GTGGGTTGT TGAAGT TGATGG TCACC	P 2	23	262,233	RF 2	4	3	1	–	21%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS159	ACCCAG GAAGAG GTGAGG TCAAGT CA	CCCAGATAG GACAGT GATAGT ACCGA	P 2	3571	60,830	RF 2	4	4	0	–	9%	0%
CS161	TTTGTTTAG GATCCA TAGCCA ACCC	GATCATTTG GCAGCT ACAGTC TGGG	P 1	1273	96,329	RF 2	7	6	1	–	8%	0%
CS163	TGGGCC GAATTG AGCTGC AAAGTGC	TCAATGTCA TGCGCT CTGTGT CAAT	P 1	4165	51,846	FL 2	22	6	4	17	6%	1.9%
CS165	ACGTAGGAT GAACAT GTCCAT TCCCA	ACACTCATC TCATCA TCCATG TTTCCT	P 1	2707	34,036	FL 2	8	7	1	1	26%	0%
CS175	GCCAGTAAA CATTGT CACCAC AACA	ACTTTCCTT TGATTC AGAGCT CTCA	P 2	6504	14,523	RF 1	3	3	0	–	15%	0%
CS176	TCTGCGTCA TCAACT CTCCAA ATGCA	CTAATGTTC CCAGGT CGCCAA GTGG	P 3	1475	106,510	RF 2	9	7	2	–	6%	2.0%
CS177	TGCTACGAT ACAACA AGACTT AAGGCA	AACTAG GAAGCG AATGTA CTTCACA	P 2	1043	147,891	FL 1	6	5	2	2	6%	0%
CS178	ACACAG GAAATG ACCCAA TAGGAGA	TCAAAC CAAGTG GGAAAG CAGAACT	P 2	147	217,131	RF 2	14	10	1	–	6%	0%
CS179	CATTCCAGC TCCAAA GTATAC AGATTC	CCACAG TGGCTT GTTTCA ACAGGT GC	P 2	4637	18,364	FL 2	11	5	2	3	6%	0%

Table 1 (continued)

Marker name	Forward sequence	Reverse sequence	Multiplex PCR	*C. mollissima* contig ID (v4.1)	Position on the contig	Genotyping strategy[a]	Number of haplotypes	Repeat variation	SNP or INDEL		Missing rate (calculated on 106 genotypes)	Genotyping error rate (calculated on 27 genotypes)
									SSR motif	Flanking seq		
CS180	AGTTCAATCCTTGACTCTCCAACCCT	TATCCAAAGACTCCGAACCAGGCTT	P 1	1435	17,889	FL 2	5	2	1	2	7%	0%
CS182	AGCGTGTTTCTTGAACCTTGCCACA	ACGCATCGTTTCTCTGCCATTCTTCA	P 3	501	179,025	RF 2	12	12	0	–	2%	0%
CS183	AGCCGCATTTGCAGAAACAACCATC	TGGTGGTGTAAGGTTCAAGACAGGA	P 2	262	174,376	RF 2	3	3	0	–	8%	0%
CS184	GCGTGATGCAAATTGGAGGCTGTGC	GAGAAACAAGATCCGGATGCACCCT	P 3	21	170,253	FL 2	6	3	2	3	5%	0%
CS186	TGCAGGGCATTCGAATGAAGAAGGT	CCTGGAAATACCCTTCGGCTTAGCT	P 1	2102	32,943	FL 2	5	4	0	2	7%	0%
CS188	TTCATGCACCACTCCTCGTCAACCA	ACGAGGGTGAATAACAATCTTGAGCGC	P 3	384	127,354	FL 2	2	2	0	1	3%	0%
CS191	ACTCATGGAGGTCGCAACTGTGGAG	GAGGGCAAGAAAGCAAGCACTCGGT	P 2	3662	18,196	FL 2	11	7	1	3	6%	0%
CS192	TGCTCTTGCCATTGCCTTCAATCCT	ACAAGCCACTGAAATTGAGGAGAGGGA	P 3	747	86,594	RF 1	3	3	0	–	5%	0%
Total/mean							887	644	112	148	8.6%	0.5%
Polymorphism partition								71.2%	12.4%	16.4%		

[a]The strategy (Repeat focus or Full length) and parameter set (1 or 2) used for genotyping (Online Resource 2)

regions and with primers located at least 20 nucleotides away from the repeat motif were kept (n = 2307) (Meglécz et al. 2014). We selected a final set of 192 primer pairs, favoring tri-nucleotide repeats, loci with more than seven repeated motifs and with known position on the *Quercus robur* reference genome (version Qrob_PM1N, Plomion et al. 2018) identified after blastn (v2.6.0, Altschul et al. 1990). Illumina specific tags were added to the 5′ end of the primer sequences: 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′ for the forward primer and 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3′ for the reverse primer. Simplex PCR amplification (Supplementary Material) was performed on genomic DNA from *C. crenata* and *C. sativa* to test the designed primers, resulting in a subset of 150 loci showing robust amplification (Online Resource 1).

The 150 primer pairs were pooled in three multiplexed PCRs according to their affinity as predicted by Primer Pooler v1.61 (Brown et al. 2017, Table 1). They were amplified (Online Resource 2) on 106 individuals, including 27 which were duplicated or triplicated, belonging to *C. mollissima*, *C. crenata* and *C. sativa* and hybrids of these species (Online Resource 3). Amplicons from the three multiplexed PCRs were pooled for each individual and a second PCR was used to add the Illumina sequencing adaptors and barcodes (Online Resource 2). Amplicons were pooled and purified using homemade Solid Phase Reversible Immobilisation beads, quantified on a LC480 II qPCR (Roche Diagnostics) using KAPA Library Quantification (Kit Roche Sequencing Solutions), size-estimated on a TapeStation 4200 (D1000 ScreenTape Assay, Agilent technologies) and sequenced using an Illumina MiSeq Reagent Kit v2 (2 × 250 bp). It produced 6,784,525 paired reads, of which 4,683,658 were kept after length filtration (> 70 nt) using Cutadapt v1.14 (Martin 2011) and after read pair merging using Pear v0.9.10 (Zhang et al. 2014, Online Resource 3). Markers were analyzed using FDSTools (Hoogenboom et al. 2017) as described previously (Lepais et al. 2020), taking into account either all polymorphisms identified in the amplicon (full length, or "FL" strategy) or only in the repeat motif sequences (repeat focus, or "RF" strategy), using different parameters for FDSTools (Online Resource 2). Loci with a genotyping error rate < 5% and with < 30% of missing data were kept for subsequent analyses. Online Resources and reads are available at https://doi.org/10.15454/PNBEAM.

Across the 150 sequenced loci, 98 markers showed consistent amplification, polymorphism and reliable genotyping with a mean genotyping error rate of 0.5% and missing data of 8.6% (Table 1). A total of 887 haplotypes were identified with an average of 9.05 haplotypes per locus (Table 1). Considering variation at the number of microsatellite motif only resulted in a total of 644 alleles with a mean number of 6.57 alleles per locus, pointing to a substantial gain when considering other sources of variation. In fact, other sequence polymorphisms (SNPs and INDEL) were detected in 76% of the loci, either within the repeated motif sequence or in the flanking sequence (Table 1). Overall, we observed 71%, 12% and 16% of the variation corresponding to microsatellite variation (repeat number), SNP or INDEL within the repeat motif, and SNP or INDEL in the flanking sequence, respectively (Table 1). Out of the tested pool, 69 markers (70%) could be located on the 12 *Quercus robur* chromosome pairs (Online Resource 1).

Acknowledgements BL was financially supported by the Region Nouvelle-Aquitaine (Grant Number 2018-1R20204 project DIGIE "chestnut DIeback: vulnerability and GenetIc determinism of ink disEase resistance). Genotyping by sequencing was conducted at the Genome Transcriptome Facility of Bordeaux (Grants from ANR-10-EQPX-16).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Barreneche T, Botta R, Robin C (2019) Advances in breeding of chestnuts. In: Serdar Ü, Fulbright D (eds) Achieving sustainable cultivation of tree nuts. Burleigh Dodds Science Publishing, Cambridge

Brown SS, Chen Y-W, Wang M et al (2017) PrimerPooler: automated primer pooling to prepare library for targeted sequencing. Biol Methods Protoc. https://doi.org/10.1093/biomethods/bpx006

Desprez-Loustau M-L, Robin C et al (2007) The fungal dimension of biological invasions. Trends Ecol Evol 22:472–480. https://doi.org/10.1016/J.TREE.2007.04.005

Guichoux E, Lagache L, Wagner S et al (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11:591–611. https://doi.org/10.1111/j.1755-0998.2011.03014.x

Hoogenboom J, van der Gaag KJ, de Leeuw RH et al (2017) FDSTools: a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. Forensic Sci Int Genet 27:27–40. https://doi.org/10.1016/j.fsigen.2016.11.007

Lepais O, Bacles CFE (2011) Comparison of random and SSR-enriched shotgun pyrosequencing for microsatellite discovery and single multiplex PCR optimization in *Acacia harpophylla* F. Muell Ex Benth Mol Ecol Resour 11:711–724. https://doi.org/10.1111/j.1755-0998.2011.03002.x

Lepais O, Chancerel E, Boury C, Salin F, Manicki A, Taillebois L, Dutech C, Aissi A, Bacles CFE, Daverat F, Launey S, Guichoux E (2020) Fast sequence-based microsatellite genotyping development workflow. PeerJ 8:e9085. https://doi.org/10.7717/peerj.9085

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10. https://doi.org/10.14806/ej.17.1.200

Meglécz E, Pech N, Gilles A et al (2014) QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. Mol Ecol Resour 14:1302–1313. https://doi.org/10.1111/1755-0998.12271

Kremer A, Abbott AG, Carlson JE et al (2012) Genomics of fagaceae. Tree Genet Genomes 8:583–810

Pereira-Lorenzo S, Ballester A, Corredoira E et al (2012) Chestnut. Fruit breeding. Springer, New York, pp 729–769

Plomion C, Aury JM, Amselem J et al (2018) Oak genome reveals facets of long lifespan. Nat Plants 4:440–452. https://doi.org/10.1038/s41477-018-0172-3

Powell WA, Newhouse AE, Coffey V (2019) Developing blight-tolerant American chestnut trees. Cold Spring Harb Perspect Biol 11:a034587. https://doi.org/10.1101/cshperspect.a034587

Staton M, Addo-Quaye C, Cannon N et al (2019) The Chinese chestnut genome: a reference for species restoration. BioRxiv. https://doi.org/10.1101/615047

The Plant List (2013). Version 1.1. Published on the Internet; https://www.theplantlist.org/ Accessed Jan 1.

Vartia S, Villanueva-Cañas JL, Finarelli J et al (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. R Soc Open Sci. https://doi.org/10.1098/rsos.150565

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620. https://doi.org/10.1093/bioinformatics/btt593