the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) doi: 10.1111/tpj.15099

RESOURCE

A chromosome-scale genome assembly of European hazel (*Corylus avellana* L.) reveals targets for crop improvement

Stuart J. Lucas[1,*], Kadriye Kahraman[1,2], Bihter Avşar[1,2], Richard J.A. Buggs[3,4] and Ipek Bilge[1,2]

[1]*Sabanci University SUNUM Nanotechnology Research and Application Centre, Istanbul, Turkey,*

[2]*Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey,*

[3]*Jodrell Laboratory, Royal Botanic Gardens, Kew, London, UK, and*

[4]*School of Biological and Chemical Sciences, Queen Mary University of London, London, UK*

Received 30 October 2019; accepted 9 November 2020.

*For correspondence (e-mail slucas@sabanciuniv.edu).

SUMMARY

The European hazelnut (*Corylus avellana* L.) is a tree crop of economic importance worldwide, but especially for northern Turkey, where the majority of production takes place. Hazelnut production is currently challenged by environmental stresses, such as a recent outbreak of severe powdery mildew disease; furthermore, allergy to hazelnuts is an increasing health concern in some regions. In order to provide a foundation for using the available hazelnut genetic resources for crop improvement, we produced a fully assembled genome sequence and annotation for a hazelnut species, from *C. avellana* cv. 'Tombul', one of the most important Turkish varieties. A hybrid sequencing strategy, combining short reads, long reads and proximity ligation methods, enabled us to resolve heterozygous regions and produce a high-quality 370-Mb assembly that agrees closely with cytogenetic studies and genetic maps of the 11 *C. avellana* chromosomes, and covers 97.8% of the estimated genome size. The genome includes 27 270 high-confidence protein-coding genes, over 20 000 of which were functionally annotated based on homology with known plant proteins. We focused particularly on gene families encoding hazelnut allergens, and the Mildew resistance Locus O (MLO) proteins that are an important susceptibility factor for powdery mildew. The complete assembly enabled us to differentiate between members of these families and to identify homologues that may be important in mildew disease and hazelnut allergy. These findings provide examples of how the genome can be used to guide research and to develop effective strategies for crop improvement in *C. avellana*.

Keywords: European hazel, *Corylus avellana*, genome assembly, powdery mildew disease, food allergy.

INTRODUCTION

The genus *Corylus* describes hazels, deciduous trees and large shrubs that are widespread throughout the northern hemisphere and are grown for their edible nuts, wood and ornamental purposes. The most economically significant species is the European hazel (*Corylus avellana* L.), the nuts of which are known as hazelnuts, filberts or cobnuts, and are consumed worldwide both directly and as an ingredient in many food and confectionary products. Hazelnuts prefer a mild, damp climate; historically, production is concentrated in the Black Sea region of Turkey, which provided approximately 65% of the world's supply in 2017 (FAO, 2017). Other major producers include Italy, Azerbaijan and the USA, and in recent years several other countries have begun actively developing their hazelnut industry, such as China, Georgia, Iran and Chile.

In spite of its widespread use, the genetic improvement of *C. avellana* as a crop has been largely limited to the American Pacific Northwest, where the devastating fungal disease Eastern filbert blight prompted a successful effort to identify and breed for genetic sources of disease resistance (Molnar and Capik, 2012; Sathuvalli *et al.*, 2017). In Turkey and elsewhere, hazelnut production is severely affected by abiotic stresses such as frost or drought, and by emerging phytopathogens such as *Erysiphe corylacearum* (Ustaoğlu, 2012; Sezer *et al.*, 2017). Over the last 3–5 years, this powdery mildew fungus has become ubiquitous in orchards in Turkey and Georgia, and controlling the disease requires repeated and costly fungicide

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd

spraying. Therefore, sources of genetic resistance to powdery mildew are urgently required. In other crop species, including *Hordeum vulgare* (barley), *Pisum sativum* (pea), *Solanum lycopersicum* (tomato), *Triticum aestivum* (wheat) and *Vitis vinifera* (grapevine), the knock-down/knockout of susceptibility genes belonging to the *Mildew Locus O* (*MLO*) family has been shown to confer resistance to powdery mildew fungi (Acevedo-Garcia *et al.*, 2014). The identification of the presence and copy number of *MLO* homologue and paralogue genes in *C. avellana* could be a first step towards developing resistant cultivars.

In recent years, nut allergies have become a well-known health problem for a minority of consumers, leading to great interest in the identification of hazelnut allergens and the genes responsible for their synthesis (Costa *et al.*, 2015). To date 11 different allergens, denoted 'Cor a' proteins, have been identified and cloned from *C. avellana*. These proteins have diverse structure and functions and some, such as Cor a 1, are found in multiple isoforms with varying levels of allergenicity (Lüttkopf *et al.*, 2002). The characterization of the genomic loci from which these proteins originate would be an important step to understanding how they are produced *in vivo*, and a foundation for developing novel and sensitive DNA-based detection methods for these allergens.

As with many tree species, *C. avellana* has a long generation time (up to 8 years to reach full productivity) and also displays sporophytic self-incompatibility, with genetically similar individuals unable to pollinate each other (Marinoni *et al.*, 2009). These factors make selecting for many important traits by classical breeding approaches extremely difficult. Therefore, genomic data, which allows the identification and selection of many genetic loci simultaneously, has huge potential to accelerate the research and breeding for *C. avellana*.

A draft genome assembly and transcriptome of the American cultivar 'Jefferson', along with resequencing data from seven further cultivars, have been produced and made publicly available (Rowley *et al.*, 2012; Rowley *et al.*, 2018). Transcriptome sequences have also been produced for two wild hazelnut species, *Corylus heterophylla* Fisch. and *Corylus mandshurica* (Ma *et al.*, 2013; Chen *et al.*, 2014). In cultivated *C. avellana*, a genetic linkage map has also been developed (Mehlenbacher *et al.*, 2006), which has been improved by the addition of simple sequence repeat (SSR) markers developed from both enrichment libraries and the available genome and transcriptome data (Gürcan *et al.*, 2010; Gürcan and Mehlenbacher, 2010; Colburn *et al.*, 2017; Bhattarai and Mehlenbacher, 2017). Recently, genotyping by sequencing has been used to generate thousands of SNP markers for a cross between two European cultivars (Tonda Gentile della Langhe × Merveille di Bollwiller), enabling the first genetic mapping of a quantitative trait – time of leaf budburst – in *C. avellana* (Marinoni *et al.*, 2018). Also, using a partial genome sequencing approach, SSR markers have been developed and used to characterize genetic diversity between Turkish and European hazelnut varieties (Öztürk *et al.*, 2018).

Although the studies mentioned above provide essential resources for the identification of genes and molecular markers in hazelnut, there is still a need for a reference quality genome sequence of *C. avellana*, in order to identify structural relationships between genes and to facilitate the rapid mapping of candidate genes from molecular markers for traits of interest. In this study, using the Turkish cultivar 'Tombul', we apply a hybrid next-generation sequencing strategy combining short-read, long-read and physical proximity sequencing to generate a *de novo* chromosome-scale genome assembly consisting of 11 pseudomolecules with a total length of 370 Mb. These pseudomolecules are compared and found to be highly consistent with previous cytogenetic data and genetic maps of the *C. avellana* genome, indicating that they represent a near-complete genome sequence. We also produce a full annotation of the genome sequence, with a detailed analysis of genes and other functional elements predicted to be involved in disease resistance and the production of hazelnut allergens.

RESULTS

A hybrid sequencing approach facilitates the complete assembly of the hazelnut genome

For an initial survey of the 'Tombul' hazelnut genome, we obtained high-coverage Illumina 150-bp paired-end reads for their low error rate and cost-effectiveness. As previously reported (Rowley *et al.*, 2018), this allowed us to produce a *de novo* draft genome assembly. This assembly was highly fragmented and 35% larger than previous estimates of the *C. avellana* genome size (378 Mb, calculated from flow cytometry data), however. This could be explained by the heterozygous regions of the genome being assembled twice into separate contigs; accordingly, analysis with BENCHMARKING UNIVERSAL SINGLE-COPY ORTHOLOGS (BUSCO 3; Waterhouse *et al.*, 2018) found that 25% of highly conserved single-copy genes from land plants (360/1440) were duplicated in the Illumina-only assembly.

Therefore, we improved the genome assembly by incorporating low-coverage, long, single-molecule reads (Oxford Nanopore Technologies, https://nanoporetech.com) and information about physically adjacent sequences produced using proximity ligation sequencing (Dovetail Genomics, https://dovetailgenomics.com). These two approaches were combined with the Illumina data separately and together, in order to assess their relative contributions to the final assembly (Table S4). Although the NanoPore reads were at relatively low genome coverage

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

(9.3×) and have a high base error rate, when combined with the Illumina data they produced a much more contiguous assembly: N50 = 78.8 kb and L50 = 1299, respectively, compared with 7.3 kb and 20 927 for Illumina only. The hybrid assembly also eliminated duplicated sequences and the majority of gaps in assembled scaffolds, giving a final assembly size of 383 Mb, very close to the expected genome size. This was still too fragmented to allow large-scale structural comparisons, for example with hazelnut genetic maps and other genomes from other species, however.

Proximity ligation sequencing by the Dovetail™ method generates pairs of linked reads that originate from within the same large DNA fragment (Chicago™ library) or from physically adjacent nucleosomes in native chromatin (HiC™ library). The sequences of the linked reads are not assembled directly but are mapped to scaffolds from a pre-existing genome assembly. As shown in Table S4, the incorporation of the Dovetail™ data with the Illumina sequences enabled the majority of the genome to be assembled into pseudomolecules of longer than 10 Mb; however, tens of thousands of small scaffolds remained unassembled, and the duplicate scaffolds were not resolved. In contrast, the Illumina+NP+Dovetail™ assembly consisted of 11 chromosome-sized pseudomolecules ranging from 22.42 to 50.95 Mb in length, in total accounting for 97.8% of the predicted genome size; the remaining unplaced scaffolds were 1 to 400 kb in size and had a total length of 13.6 Mb. The chromosome-sized pseudomolecules (hereafter 'chromosomes' for brevity) were labelled pchr01–pchr11 in descending order of size. The completeness of the assembly was confirmed by BUSCO analysis: orthologues of 97% of 1440 highly conserved land plant genes were found in the chromosomes (90% complete single copies, 6% complete and duplicated, and 1% fragmented).

We assessed the large-scale accuracy of the hybrid assembly by comparison with previously published cytogenetic and genetic maps of the *C. avellana* chromosomes. Falistocco and Marconi confirmed the karyotype of diploid *C. avellana* as $2n = 22$ and noted that there were three distinct size groups of two large, five medium and four small chromosomes (Falistocco and Marconi, 2013). Similarly, our hybrid assembly contains 11 chromosomes that fall into the same three size groups (Table S5). The aforementioned study also used *in situ* hybridization to locate *45S* and *5S* ribosomal DNA (rDNA) repeats on one chromosome from the large and small groups, respectively; using BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST), we located a *45S* rDNA on pchr02 and a *5S* rDNA on pchr11.

Furthermore, known *C. avellana* SSR sequences were mapped onto the pseudomolecules and compared with the genetic map of the cross OSU 252.146 × OSU 414.062 (Colburn *et al.*, 2017) (Figure S1). Each chromosome only contained SSRs from a single linkage group, meaning that they could be unambiguously identified, and 108/113 (95.6%) of the shared SSRs were co-linear in the two data sets, suggesting that there are no large-scale structural differences between the 'Tombul' chromosomes and the earlier genetic map.

Taken together, these data indicate that the chromosome sequences presented here are consistent with what is known about the physical structure of the hazelnut genome, and that the few differences detected may result from local translocations specific to the 'Tombul' variety. Therefore, we propose that these data can be used as a reference genome for continuing studies, especially for Turkish hazelnut varieties.

Functional annotation of the hazelnut genome

The chromosome sequences were annotated using a combination of transcriptome sequencing and sequence homology with previously reported genetic elements, as described below and in Appendix S1. Repetitive sequences and genes encoding microRNA (miRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA), and proteins, were mapped to all chromosomes (Figure 1). As is typical for eukaryotic genomes, transcribed genes were more abundant towards the ends of the chromosomes, whereas repetitive DNA content was concentrated near the centromeres. Protein-coding genes were also assigned to orthologue groups (OGs) using ORTHOMCL (Fischer *et al.*, 2011). It was observed that the large number of gene models (4324) were adjacent to other genes from the same OG, suggesting that these genes are paralogues resulting from local tandem duplications. These tandem duplicated genes were also generally more abundant near the ends of the chromosome arms, and often coincided with regions of highest gene density (Figure 1, innermost two tracks).

Hazelnut transposable elements are largely degenerate, with little recent activity

An initial screening of the genome assembly with repetitive elements previously annotated in other eudicots detected few matches (11.87% of the genome), suggesting that the majority of repetitive elements in the genome are lineage specific. Therefore, prediction tools were used to generate a database of *Corylus*-specific transposable elements (TEs) based on known structural features of each repeat class (Appendix S1; Table S1). When these were included, 35.72% of the entire genome assembly was found to consist of interspersed repeats, whereas a further 2.41% was made up of simple repeats and low-complexity sequences (Appendix S1; Table S7). Repetitive content varied widely along each chromosome, from ≤25% near the ends to 75–90% in the pericentromeric regions. Over 92.7% of the repetitive DNA comprised retroelements with long terminal repeats (LTRs). Over half of these elements were

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

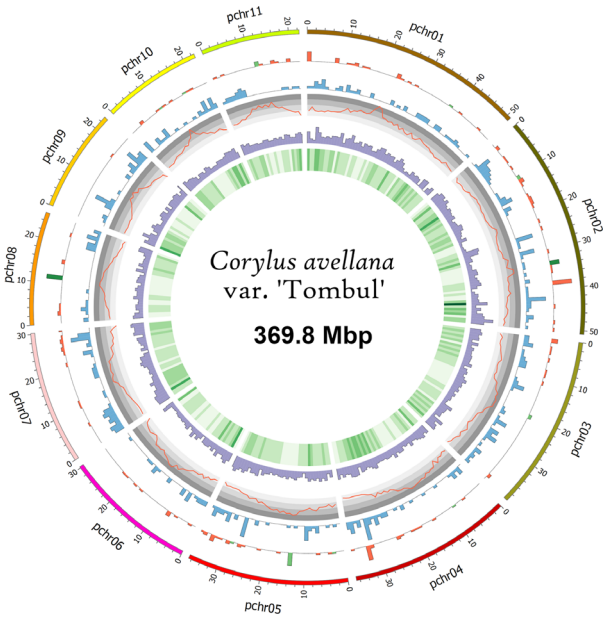


Figure 1. Circular plot of the *Corylus avellana* cv. Tombul genome summarizing the functional features (detailed in Tables S7–S13). Working in from the outside: ideogram of pseudochromosomes, with lengths marked in Mbp; histogram of miRNA (red), *5S* rRNA (green) and *45S* rRNA (dark green) gene density; histogram of tRNA genes (blue); line graph of repetitive content as percentage of total sequence (background shading from light to dark grey indicates the interquartile ranges); histogram of protein-coding gene density (purple); heat map of tandem gene duplications (where darker green shading indicates more duplications).

degenerate, with internal deletions and too much sequence diversification to positively assign them to a repeat family. Of the remainder, Copia elements were almost twice as abundant as Gypsy elements (Appendix S1; Figure S2). In the *Betula pendula* genome, some classes of terminal inverted repeat (TIR) transposons and terminal-repeat retrotransposons in miniature (TRIM) retroelements were highly abundant (Salojärvi *et al.*, 2017). This was not the case in *C. avellana*, however, indicating that the expansion of these families took place only in the *Betula* lineage.

These observations suggest that TEs have been relatively static in the recent history of *Corylus*, with the majority permanently inactivated by mutations and deletions. This is in contrast to *Betula* and to many other plant species, where TEs are cited as a primary driver for rapid genetic diversification (Niu *et al.*, 2019).

Corylus avellana microRNAs are regulators of development and stress responses

Genes coding for functional non-coding RNAs were predicted and annotated as described below. A total of 477 predicted tRNA genes and 40 tRNA pseudogenes were distributed across all of the chromosomes (Figure 1; Table S8), representing all 20 amino acids and 54 of the possible anticodons. These included 10 putative suppressor tRNAs, with anticodons complementary to the TGA (9) or TAA (1) stop codons. tRNA types and their codon preferences were also analysed in three other tree species (Appendix S1; Figure S3; Table S9). Ribosomal RNA genes were found on pchr02 and pchr08 (*45S* rDNA) and pchr05 and pchr11 (*5S* rDNA), whereas ribosomal proteins were distributed among all chromosomes except pchr11 (Table S10).

MicroRNAs are ubiquitous post-transcriptional regulators in plants and a previous study identified putative miRNA genes in the draft *C. avellana* cv. Jefferson genome (Avsar and Aliabadi, 2017). In the Tombul genome, 153 putative conserved miRNA genes were annotated, including members of 52 different miRNA families (Table S11).

The majority (95/153) of predicted mature miRNA sequences were 21 nt in length, and the most abundant miRNAs were the well-characterized miR156, miR171 and miR399 families, with 16, 12 and 12 candidates, respectively (Appendix S1; Figure S4). Mapping the predicted pre-miRNAs to the transcriptome sequences assembled from leaf buds found evidence for the expression of 22/52 miRNA families in this tissue, many of which have also been functionally annotated in other species (Table 1). Other potential targets for the miRNAs identified in this study were predicted by searching for miRNA complementary sequences in unspliced transcript sequences of the high-confidence *C. avellana* gene models (see below). Targets were identified for all predicted mature miRNAs except one (an miR395 variant) and 10 852 (39.8%) of these genes were predicted to be regulated by at least one miRNA (Table S12).

The 52 conserved miRNA families found in *C. avellana* cv. Tombul were also compared with those predicted by the same method from the draft cv. Jefferson genome and five other tree species (Figure 2a). Although there was a well-conserved group of 26 miRNA families common to all of the species examined, there were also miRNAs that were unique to each species. Surprisingly, there were eight miRNA candidates that were predicted in Tombul but not in Jefferson, and four for which the reverse was true. Although there were no transcripts for these miRNAs in our data set, two of them (miR1520 and miR7486) were recently identified by small RNA sequencing from dried hazelnuts (Aquilano *et al.*, 2019), suggesting that these miRNAs are expressed only in specific tissues or in response to environmental stimuli. Also of interest are miR1863 and miR8738, which were identified in both Tombul and Jefferson but not in other tree species, and were supported by transcripts; miR8738 was also recovered from dried nuts. These miRNAs may play roles in hazelnut-specific gene regulation networks.

The set of candidate target genes for all miRNAs were annotated with gene ontology (GO) terms, and enrichment analysis was carried out using the whole gene set as a

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

Table 1 miRNA families expressed in *Corylus avellana* buds under normal conditions

miRNA family (loci in genome)	Targets confirmed in other species	Reported biological functions
miR156/157 (16)	Squamosa-promoter binding protein (SBP) box transcription factors	Shoot and leaf development
miR160 (2)	Auxin response factor (ARF) proteins	Auxin signalling, plant development
miR162 (1)	Dicer-like 1 (DCL1) proteins	miRNA processing
miR164 (3)	NAC domain transcription factors	Shoot apical meristem formation, abiotic stress
miR165/166 (4)	HD-Zip transcription factors	Meristem and leaf development
miR167 (5)	Auxin response factor (ARF) proteins	Floral development, stress responses
miR169 (8)	CCAAT-box binding/NF-Y transcription factors	Abiotic and biotic stress responses
miR170/171 (12)	GRAS-domain or SCARECROW-like transcription factors	Root patterning, light and gibberellin signalling
miR172 (4)	APETALA2-like transcription factors	Plant development
miR319 (4)	TCP-like transcription factors	Leaf development, abiotic stress response
miR393 (1)	F-box proteins & bHLH transcription factors	Root development, auxin signalling
miR394 (4)	F-box proteins	Plant development, stress responses
miR396 (3)	GRF transcription factors, rhodenase-like proteins, kinesin-like protein B	Growth regulation
miR403 (3), miR482 (5), miR529 (1), miR1863 (2), miR5021 (7), miR6288 (1), miR8738 (1), miR9752 (1)		*No experimentally confirmed targets*

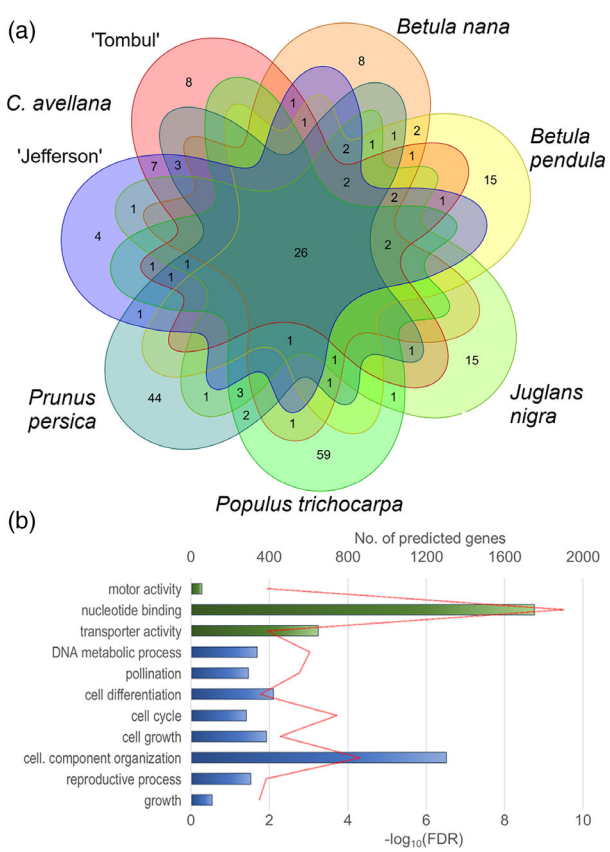


Figure 2. (a) Venn diagram of conserved plant miRNA families identified in *Corylus avellana* cv. Tombul and other published tree genomes. (b) Gene ontology (GO) terms significantly enriched in predicted miRNA target genes. Bars show number of target genes for each GO term in the molecular function (MF, green) and biological process (BP, blue) domains; the red line shows the negative $\log_{10}$ of false discovery rate (FDR)-corrected *P* values.

reference. This indicated several GO terms that were significantly enriched among miRNA targets (Figure 2b). The most abundant and significantly enriched molecular function (MF) term was 'nucleotide binding', a term that encompasses many proteins involved in intracellular signalling, such as kinases, receptor-coupled G proteins and cyclic AMP binding proteins.

The significantly enriched biological process (BP) terms included cellular component organization, cell differentiation, cell growth and related processes. Taken together, these results suggest that although individual miRNAs regulate diverse functions, at a system level the miRNA mechanism in hazelnut is involved in the control of plant growth and development in response to external stimuli. Therefore, miRNA regulation would be a valuable subject for research into stress and environmental adaptation in *Corylus* and other nut crops.

It has also been demonstrated that some nut miRNAs can interact with genes from the mammalian immune system, such as miR-156c with the tumour necrosis factor α (TNF-α) receptor (Aquilano *et al.*, 2019). In light of the known allergenicity of hazelnuts, further investigation of *Corylus*-specific nut miRNAs would also be of great interest.

The *C. avellana* cv. Tombul genome was formed by ancient segmental duplications and recent gene family expansion

Using an *ab initio* method, 50 761 gene models were predicted in the repeat-masked chromosomes. These were compared with OGs of proteins from well-characterized plant genomes. Of our models, 12 516 (25%) were not found in any other plant species searched. Gene models

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

were considered to be of low confidence if they did not have orthologues in other species, and if they did not have BLAST matches in the Tombul bud transcriptome assembly or a match to a paralogue that did match a transcript (for thresholds, see Appendix S1). This resulted in a catalogue of 27 270 high-confidence protein-coding gene models (Table S13), 16 521 of which were supported by high transcript coverage under normal growth conditions. There were 2844 high-confidence gene models that were expressed in leaf buds but had no homologues outside of hazelnut, suggesting that these are *Corylus*-specific genes. During automated annotation, InterPro protein domains were identified in 88.77% of gene models (24 208), whereas 82.38% of gene models (22 466) were identified with one or more GO terms, and 49.07% (13 381) were assigned to MAPMAN functional bins representing plant-specific molecular components and pathways (Tables S13 and S14). Only 2332 gene models (8.55%) were not assigned functional annotation by any of these approaches.

The comparison of complete gene sets from different genomes makes it possible to select large numbers of highly conserved single-copy genes for phylogeny building, and also to identify patterns of gene presence/absence and duplication in particular lineages. With this in mind, the high-confidence *C. avellana* gene set defined above was compared with those from the other available high-quality genomes from the Fagales (*B. pendula, Juglans regia* and *Quercus robur*) and two genomes from the Rosaceae (*Malus domestica* and *Prunus persica*) as an outgroup, using ORTHOFINDER to identify OGs and reconstruct their phylogenetic relationships (see Experimental procedures). The large majority of *C. avellana* genes (86.3%) were assigned to cross-species OGs, with the remainder either assigned to *Corylus*-specific OGs or having singleton status within *Corylus* (6.8 and 6.9%, respectively). A total of 979 OGs contained a single gene from each of the six species genomes; sequence comparisons between these single-copy orthologues were concatenated and used to build a dated phylogeny using a relaxed molecular clock (Dos Reis *et al.*, 2016). The divergence between the ancestors of *Betula* and *Corylus* was estimated to have occurred 51 Mya (95% confidence interval 45.2–62.8 Mya; Figure 3a), which is a more recent date than reported from previous estimates (63.5–78.0 Mya; Larson-Johnson, 2016) inferred from alignments of the *ITS* region and four organellar genes (*matK*, *matR*, *rbcL* and *trnL*). The date estimates for other nodes of the tree all overlapped with

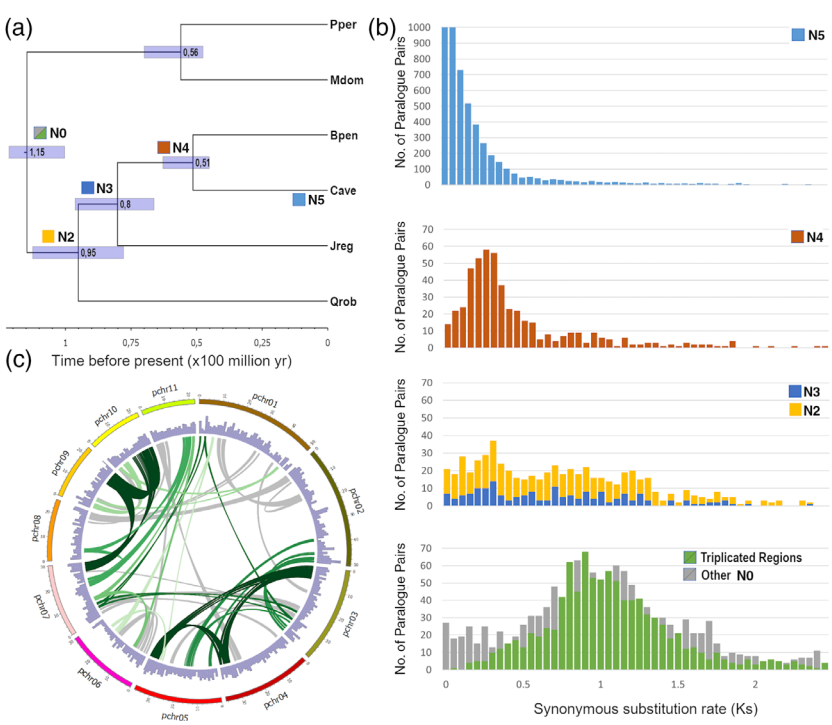


Figure 3. (a) Fossil-calibrated phylogenetic tree generated using single-copy orthologues shared between all species (Bpen, *Betula pendula*; Cave, *Corylus avellana*; Jreg, *Juglans regia*; Mdom, *Malus domestica*; Pper, *Prunus persica*; Qrob, *Quercus robur*). Node labels indicate the estimated time of each speciation event (×100 Mya), whereas bars give 95% confidence intervals. Coloured boxes and node IDs (N0–N5) are given as a key for the histograms. (b) Histograms showing the number and K_s distribution of paralogous gene pairs from the *C. avellana* genome originating at or before each node in the species tree. (c) Circular plot of the *C. avellana* genome with histogram of gene density, and links indicating intragenomic collinear regions consisting of 10 or more genes; triplications are shown with green links, duplications with grey.

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

those from previous studies (Magallón *et al.*, 2015; Larson-Johnson, 2016). Based on these dates, the rate of point mutations on the *Corylus* branch of the tree was calculated to be 0.14 substitutions per base every 100 Myr, approximately 50% faster than the *Betula* and *Quercus* branches and twice as fast as *Juglans*.

Paralogues resulting from gene duplications within the *Corylus* genome were also analysed. The node of the species tree at which each gene duplication occurred was identified by determining whether both members of a paralogue pair are present in >50% of genomes descended from that node. On this basis, 1917 gene duplications were indicated to have occurred before the divergence of Fagales and Rosales (Figure 3a, node N0), whereas a further 714 occurred in the common ancestor of Fagales (node N2). Continuing down the *Corylus* lineage, 280 and 699 gene duplications were indicated at nodes N3 and N4, respectively, whereas 6424 were unique to the *Corylus* genome (N5).

Pairs of paralogues resulting from each gene duplication event were aligned with each other using CLUSTALW, and substitution rates were calculated. The rate of synonymous substitutions in protein-coding sequences is assumed to be independent of selection pressures and therefore the number of synonymous substitutions per site (K_s) will increase steadily over time. The distribution of K_s values for the paralogue pairs assigned to each node of the species tree were plotted (Figure 3b) and, as expected, the large majority of *Corylus*-specific paralogue pairs had K_s values of <0.1, suggesting that they are the product of recent gene duplications (node N5, top panel). The K_s distribution for paralogue pairs assigned to earlier nodes on the time tree moved to progressively higher values with increasing node age. Of particular interest is the peak in the distribution for N0 paralogue pairs (K_s = 0.9–1.1), suggesting that a large number of gene duplications occurred at a similar time.

To investigate this further an all-against-all comparison of protein sequences was used to identify blocks of 10 or more collinear paralogues within the *Corylus* genome using MCSCANX (Wang *et al.*, 2012), which are indicative of past large-scale duplications. Sizeable segments of triplicate collinearity were evident, for example between the distal ends of chromosomes pchr07, pchr09 and pchr10, and the proximal end of pchr03 with the distal ends of pchr04 and pchr05 (Figure 3c, dark-green links). Significantly, cross-referencing these genes with the duplication events described above showed that the large majority of genes in collinear blocks were assigned to node N0 and coincided with the peak in K_s values (Figure 3b, bottom panel). This indicates that the triplicate blocks in the *C. avellana* genome originated before the diversification of Fagales and Rosales; this is consistent with observations from other species, indicating that a whole-genome triplication event, referred to as γ, took place in the common ancestor of the core eudicots (Jiao *et al.*, 2012). All the collinear blocks indicated in Figure 3c had average K_s values of >0.9, providing no evidence for any large segmental duplications in the more recent history of *Corylus*. The MCSCANX analysis also identified 6861 tandem duplicates, largely corresponding to the *Corylus*-specific paralogue pairs assigned to node N5. Furthermore, whole-genome alignment of the *C. avellana* pseudomolecules with those from *B. pendula* showed that the chromosome structure of the two is very highly conserved (Appendix S1; Figure S5). Taken together, these observations reveal that the large-scale structure of the *C. avellana* genome has been very stable since the diversification of the Betulaceae. In contrast, there have been many small-scale gene duplications, with approximately 23% of modern *C. avellana* genes indicated to be the product of recent gene family expansion.

Corylus-specific gene families encode adaptations to diverse stress conditions

Functional annotation of the *C. avellana* gene models using the MAPMAN ontology identified genes for all major plant BPs in the genome, with the largest categories being related to RNA transcription and protein synthesis, modification and regulation (Figure 4a). Using the OGs defined above, gene families that have expanded or contracted in the *Corylus* lineage were identified by comparing the number of OG members in *C. avellana* with the total number in the three other Fagales genomes, using the chi-square test for significance (with Benjamini–Hochberg correction; false-discovery rate, FDR < 0.05). Among OGs with six or more members in Fagales, 147 were unique to or significantly expanded in *C. avellana*, whereas 33 were significantly contracted (Figure 4b). Many of the largest *Corylus*-expanded OGs currently have no functional annotation (or are only annotated with protein domains of unknown/diverse functions), suggesting that these gene families may have functions unique to the *Corylus* lineage. Other expanded OGs included gene families known to be involved in pathogen recognition (lectins and nucleotide-binding site leucine-rich repeat genes, NBS-LRRs) and intercellular signalling (leucine-rich repeat receptor-like protein kinases, LRR-RLKs, and non-cysteine-rich peptide, NCRP, signalling peptides). Single enrichment analysis revealed that 244 GO terms were significantly enriched (FDR < 0.05) in members of the *Corylus*-expanded OGs, relative to the whole gene set (Table S15). The most informative enriched GO terms from the BP domain are summarized in Figure 4c. Protein kinase activity was the single largest enriched GO term, found in almost one in six of the *Corylus*-expanded OG members, followed by several terms related to pathogen defence responses. Also enriched were terms connected to phytohormone production, abiotic stress responses, detoxification of chemicals, development

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

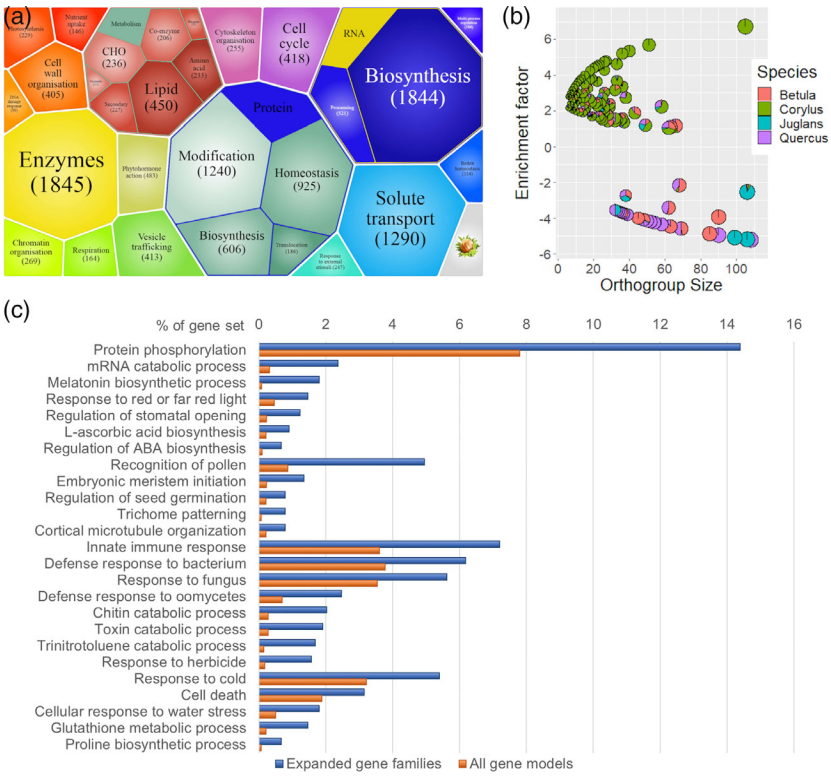


Figure 4. (a) Distribution of annotated *Corylus avellana* genes into MAPMAN bins, with the size of each polygon proportional to the number of genes assigned to the bin. (b) Scatter plot of orthologue groups (OGs) significantly enriched or reduced in *C. avellana* compared with three other Fagales genomes. The *y*-axis shows Log_2-scaled relative increase/decrease in OG size in hazelnut, compared with the average no. of OG members per genome. (c) Gene ontology (GO) terms from the biological process (BP) field found to be significantly enriched in OGs that have undergone expansion in *C. avellana*.

and pollen recognition. In summary, the gene families that are expanded specifically in *Corylus* encode predicted receptors, signalling pathway components and effectors involved in responses to a diverse range of environmental stress factors.

Inter-chromosomal homology provides insight into the evolution of the *MLO* gene family in the Fagales

The MLO proteins were first identified in barley, where a loss-of-function mutation in the gene *Mildew Resistance Locus O* conferred durable resistance to nearly all strains of the barley powdery mildew pathogen *Blumeria graminis* f.sp. *hordei* (Büschges *et al.*, 1997). Although powdery mildew is caused by different fungal species on each plant host, the resistance to mildew infection of *MLO* mutants has since been observed in *Arabidopsis thaliana*, tomato and pea (Acevedo-Garcia *et al.*, 2014), and was introduced by gene editing in wheat (Wang *et al.*, 2014). Therefore, this mechanism seems to be functionally conserved across diverse plant species and could present a promising target for developing *Erysiphe corylacearum* powdery mildew resistance in hazel.

The annotation pipeline described above identified 27 gene models that contained the InterPro domain IPR004326, 'MLO-related protein'. These were examined in detail and manually re-annotated as described in Table S16. On the basis of sequence similarity and secondary structure prediction, 12 of these gene models were found to encode full-length MLO proteins, whereas most of the remainder appeared to be truncated orthologues of the full-length genes. In order to identify the genes most likely to be involved in powdery mildew infection, the predicted protein sequences were aligned with the original MLO protein from barley (HvMLO), and with those previously described in *A. thaliana* and *Malus domestica* (apple) (Pessina *et al.*, 2014), the most closely related species to hazel in which this family has been studied in depth. A phylogenetic tree was used to cluster the sequences, which formed seven clades, in agreement with previous studies (Figures 5a and S6). No *C. avellana* MLO proteins were found in clade IV, which consists of mostly monocot mildew resistance genes, or in clade VIII, which has only been reported in some Rosaceae species (here represented by *MdMLO20*). Interestingly, the largest number of *C. avellana* *MLO* genes (four) fell into clade V, which contains all of the dicot *MLO* genes that are known to confer susceptibility to powdery mildew; of those shown here, *AtMLO2*, *AtMLO6* and *AtMLO12*, along with *MdMLO19*. Two other clades, VI and VII, link clade IV with clade V in the phylogeny. The local genome context and OGs containing the

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

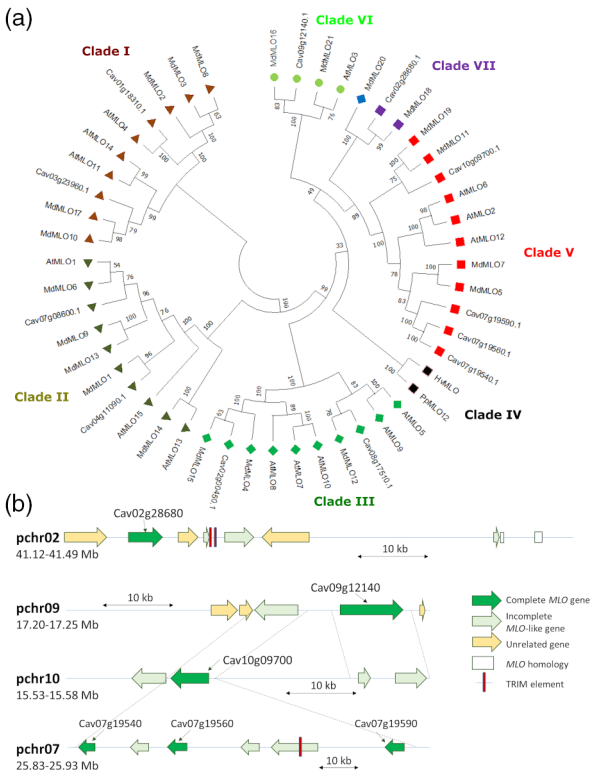


Figure 5. (a) Phylogenetic clustering of *Corylus avellana MLO* gene models with those from *Arabidopsis thaliana* and *Malus domestica*, using the maximum-likelihood method with variable rates between sites; node confidence values are a percentage of 1000 bootstrap replications. (b) Schematic of gene features in chromosome regions containing *MLO* gene models from clades V, VI and VII. Dotted lines indicate paralogous gene relationships between pchr07, pchr09 and pchr10.

C. avellana MLO genes were also examined in detail (Figures 5b and S7). As expected from the phylogeny, on pairwise comparison most of the full-length *MLO* genes had K_s values of >1, indicating that the diversification of these genes took place in the common ancestor of the eudicots. Exceptions were observed, however: in the case of *Cav02g28680*, two disrupted *MLO* genes were found within 50 kb. In the first, an insertion of two TRIM elements near the beginning of the gene split it into two shorter open reading frames. In the second, a stop codon truncated the protein after the first 180 amino acids, but the remaining exon sequences were still present downstream (Figure 5b, top). Comparison with other tree genomes showed that these truncated *MLO* genes are the result of gene duplications that are unique to *Corylus* (Figure S7a).

A more complex history could be reconstructed for the clade-V and -VII *MLO* genes, which fall within the larger region of genes showing collinearity between chromosomes 7, 9 and 10. *Cav9g12140* was found back-to-back with a second *MLO*-like gene on the negative strand, which is now truncated. These two genes were the closest paralogues of a similar pair of genes on pchr10, except that the gene on the negative strand (*Cav10g09700*) is intact and its partner on the positive strand is disrupted into two shorter open reading frames. *Cav10g09700* also has a truncated tandem duplicate immediately downstream and is the closest inter-chromosomal paralogue of the cluster of three *MLO* and three *MLO*-like genes on pchr07. This cluster of genes is represented by a single orthologue in *B. pendula* and the other tree genomes (Figure S7b).

Based on these observations, we propose that an ancestral locus existed in early eudicots that contained two *MLO* genes in back-to-back orientation; the γ genome triplication then gave rise to three such loci from which the modern clade-V and -VI *MLO* genes evolved. In the *Corylus* lineage only one gene remained active at each locus, whereas the second gene degenerated; on pchr09 the clade-VI gene (positive strand) and on pchr07 and pchr10 the clade-V gene (negative strand) remained intact. Furthermore, repeated tandem duplications led to *Corylus*-specific expansion and diversification of the clade-V *MLO* genes, especially on pchr07; no remnants of the posited ancestral clade-VI gene were found at this locus. Apart from *Cav2g28680* mentioned above, other hazelnut *MLO* genes did not have any recent tandem duplicates, suggesting that there has been specific selective pressure for the diversification of the clade-V *MLO* genes in the *Corylus* lineage.

Genomic insights into hazelnut allergenicity

To date, hazelnut allergens have been identified empirically by screening hazelnut protein extracts with sera from nut-allergic patients. Proteins that show specific immunoglobulin E (IgE) reactivity were then partially identified by Edman sequencing; this provides enough sequence to design primers and retrieve the complete coding sequence by reverse-transcriptase polymerase chain reaction (RT-PCR) (Beyer *et al.*, 2002; Schocker *et al.*, 2004). By this approach, 11 hazelnut allergens have been identified and recorded in the WHO allergen database so far (http://www.allergen.org), whereas a 12th (Cor a TLP) has been reported but is not yet confirmed (Palacín *et al.*, 2012). We used the published sequences of these allergens to identify their coding genes, and homologues, in our genome annotation (Table 2). We found gene models predicted to encode proteins with 96–100% amino acid identity to all the known allergens; the few variations can be attributed to sequence diversity between hazelnut cultivars.

Food allergens frequently comprise groups of closely related proteins that show IgE cross-reactivity, including across species. 'Isoallergens' are defined as proteins from the same species that show cross-reactivity and at least 67% amino acid identity. Within these, 'isoforms' are considered to be variants of the same allergen, typically with

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

Table 2 Genes encoding hazelnut allergen proteins in *Corylus avellana* cv. Tombul

Allergen name	Allergen group	Reference sequence (s)	Closest gene model(s)	Identity (%)	Genome orthologues
Cor a 1.01 (1.0101–1.0104)	Major hazelnut allergen. Stress-induced proteins of 159–161 aa,	CAA50327.1	*Cav01g11770*	98.75	*Cav01g11860*, *Cav01g11870* and *Cav01g11880* also have 88–95%
Cor a 1.02	members of the PR-10 protein	CAA50328.1	*Cav01g11790*	99.38	identity to Cor a 1.03
Cor a 1.03	family. Cross-reactive with	CAA50325.1	*Cav01g11770*	98.13	
Cor a 1.04 (1.0401–1.0404)	Bet v 1	CAA50326.1	*Cav01g11790*	98.75	
		CAA96548.1	*Cav01g11820*	96.88	
		CAA96549.1	*Cav01g11830*	96.23	A gene family with 40–50% identity
			Cav01g11900	96.23	to Cor a 1 also has eight members
		AAD48405.1			in this region, and five on other
		AAG40329.1	*Cav01g11730*	100.00	chromosomes
		AAG40330.1	*Cav01g11750*	93.79	
		AAG40331.1			
Cor a 2.01 (2.0101–2.0102)	Profilins. Actin-binding proteins of 131 aa	AAK01235.1	*Cav11g13740*	98.47	All >80% identity: *Cav07g09370*,
		AAK01236.1		98.47	*Cav11g13750*, *Cav11g13790*, *Cav11g13800*
Cor a 6	Isoflavone reductase-like	AGU09563.1	*Cav06g05890*	99.70	*Cav06g05880* (83.4% identity)
Cor a 8	nsLTP (non-specific lipid transfer protein), PR-14	AAK28533.1	*Cav09g17520*	100.00 (2-aa deletion)	Nine other putative nsLTPs were identified, all <50% identity to Cor a 8
Cor a 9	Legumin/*11S* globulin seed storage protein	AAL73404.1	*Cav11g01160*	99.0	*Cav11g01190* (74% identity)
			Cav11g01180	97.0	
Cor a 10	HSP70-family luminal protein	CAC14168.1	*Cav03g25070*	99.0	All other HSP70s <50% identity
Cor a 11	Vicilin/*7S* globulin seed storage protein	AAL86739.1	*Cav02g11550*	99.1	23 aa longer than reference. No predicted orthologues
Cor a 14	*2S* Albumin, seed storage protein	ACO56333.1	*Cav03g25590*	99.3	*Cav03g25600* is 75% identical
Cor a 12	Oleosins	AAO67349.2	*Cav05g23730*	100.0	*Cav03g05380* is 60% identical to
Cor a 13		AAO65960.1	*Cav08g11250*	99.3	Cor a 12; *Cav06g12600*
Cor a 15		MK737923.1[b]		nd[b]	
Cor a TLP	Thaumatin-like proteins, PR-5 family	Palacín *et al.* (2012)	*Cav09g00740*	88.0[a]	Of 48 TLPs in the genome, seven are >75% identical to Mal d 2

[a]GenBank accession IDs of protein sequences are given.
[b]MK737923 sequence is not public at the time of writing but, based on its size and similarity to Cor a 12, *Cav03g05380* is the best candidate gene.

>90% identity (Chapman *et al.*, 2007). The first and most fully characterized hazelnut allergen is Cor a 1, which is reported to include four isoallergens, two of which themselves have four isoforms (Breiteneder *et al.*, 1993; Lüttkopf *et al.*, 2002). Cor a 1 is also cross-reactive with paralogues from other Betulaceae species, such as the pollen allergens Bet v 1 from birch and Car b 1 from hornbeam. We found that gene models for all the previously reported Cor a 1 proteins were found within a single locus on pchr01 (from 12.3–12.7 Mb); this locus included 20 predicted Cor a 1 homologues, interspersed with unrelated genes. Twelve of the genes were >85% identical to known Cor a 1 sequences; all of these encoded proteins of 159–161 amino acids in length that had a two-exon structure with the intron interrupting codon 62, typical of the Bet v 1 allergens (Hoffmann-Sommergruber *et al.*, 1997). The remaining eight homologues, along with five others found on other chromosomes, formed a group with variable intron–exon structure and protein sizes (112–172 aa) and only 40–50% identity to Cor a 1. These were labelled as 'Cor a 1-like' proteins, none of which has been shown to have allergenic activity.

We carried out a phylogenetic reconstruction and fossil-calibrated time tree analysis of predicted Cor a 1 homologues from the available tree genomes, using the same methods as used in Figure 3a, revealing that the relationship between these gene loci and the established isoallergen nomenclature is complex (Figure 6; Table 2). Two genes encoding 160 amino acid proteins are predicted to express the four isoforms of Cor a 1.01, with Cor a 1.0101 and Cor a 1.0103 most closely matching *Cav01g11770*, whereas Cor a 1.0102 and Cor a 1.0104 matched *Cav01g11790*. We suggest that different alleles of these two genes account for the multiple reported isoform variants. *Cav01g11820* was the only close match to Cor a 1.02; in contrast, Cor a 1.03 had no single best match but had >90% identity to a cluster of six genes encoding 159 amino acid proteins, making all of these potential new isoforms of this allergen. All four reported isoforms of Cor a 1.04 appear to be allelic variants of *Cav01g11730*, which is

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

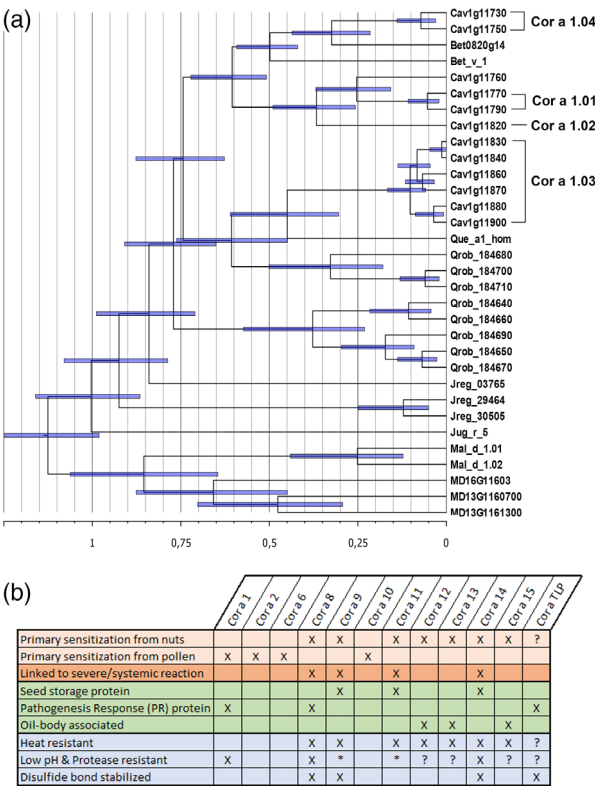


Figure 6. (a) Bet v 1/Cor a 1 homologues from Fagales were clustered by the maximum-likelihiid (ML) method with variable rates, and a fossil-calibrated time tree was calculated. Node bars give 95% confidence intervals of the divergence time, scale units are 100 Mya. (b) Shared allergenic, functional and biochemical characteristics of known and suspected hazelnut allergens. **11S* and *7S* globulins are partially digested, but leave smaller, protease-resistant polypeptides.

identical in sequence to Cor a 1.0403; however, *Cav01g11750* is also >90% identical to Cor a 1.04, and so may encode additional isoforms. In addition, *Cav01g11760* is the gene with highest similarity to Bet v 1.01 in the *C. avellana* genome, but the time tree suggests that it diverged from Cor a 1.01 and Cor a 1.02 (85% identity); therefore, it may encode a new isoallergen, putative Cor a 1.05. All new putative isoallergens and allergen-like genes are listed in Table S17.

Genomic analysis of most of the other known hazelnut allergens was much more straightforward. Cor a 6 and Cor a 8– Cor a 14 all matched single gene models with >99% identity, indicating that these are the unique genes expressing these allergens. Cor a 15 is a recently reported oleosin of similar size to Cor a 12 but with differing N- and C-terminal sequences (www.allergen.org). Although the protein sequence of Cor a 15 is not in the public domain at the time of writing, based on the available information *Cav03g05380* is the probable gene.

Cav11g01160, which encodes Cor a 9.0101, had an adjacent 97% identical duplicate (*Cav11g01170*) encoding a putative new isoform; similarly, *Cav06g05880* and *Cav03g25600* might be isoallergens of Cor a 6 (*Cav06g05890*) and Cor a 14 (*Cav03g25590*), respectively. The profilin allergen Cor a 2 has two reported isoforms, both of which were equally close matches to *Cav11g13740*; however, two new isoallergens of >80% identity to Cor a 2 were also identified (*Cav07g09370* and *Cav11g13750*), and these were identical in sequence to profilins previously isolated from hazelnut pollen (Jimenez-Lopez *et al.*, 2012). A third profilin gene, which appears to have been divided in half by a local translocation, were found in the vicinity of Cor a 2 (*Cav11g13790* and *Cav11g13800*). With all potential isoallergens, serological tests would be required to determine whether or not these homologues are allergenic.

Thaumatin-like proteins (TLPs) are known to be important causes of fruit allergy, especially in the Rosaceae. A TLP isolated from *C. avellana* was previously reported to have cross-reactivity with sera from fruit-allergic patients (Palacín *et al.*, 2012), but in <10% of cases. The GenBank accession ID given for Cor a TLP in the aforementioned article actually refers to an apple TLP; however, based on the peptide sequences also reported, we inferred that the Cor a TLP tested was encoded by *Cav09g00740*. This is one of 48 predicted TLPs found in the hazelnut genome; seven of these, including *Cav09g00740*, have >75% identity to known allergens Mal d 2 (apple) and Pru p 2 (peach). Therefore, although TLPs have not yet been demonstrated to be allergenic in hazelnut, there is high potential for cross-reactivity between these genes and homologous fruit allergens.

In summary, we identified complete gene models encoding all known hazelnut allergens and several previously unreported potential allergenic proteins, including nine new isoforms, four new isoallergens, and suspected new cross-reactive oleosin and TLP proteins (Table S17).

DISCUSSION

Towards a reference genome for *C. avellana*

Hazelnut is typical of a number of important crop species for which, until recently, limited genome data have been available. The publicly available *C. avellana* var. 'Jefferson' draft genome sequence made it possible to identify the majority of gene sequences, but not their chromosomal locations. Here, we aimed to produce a chromosome-level genome assembly for the Turkish cultivar 'Tombul' in a time- and cost-effective manner. This required three different sequencing technologies that provided data with different size ranges: 0.1–1 kbp (Illumina paired-end), 1–10 kbp (NanoPore) and 10 kbp–10 Mbp (Dovetail™). None of these methods individually or in pairs was sufficient to reconstruct the whole genome but combining all three together produced a chromosome-scale genome assembly.

The chromosomes presented here have a total length of 370 Mb, about 2.1% shorter than the estimated genome

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

size. Each chromosome still contains several hundred small sequence gaps (Table S5), the actual size of which is not known. Also, it is likely that telomeric and centromeric repeats are more condensed in our assembly than in the physical chromosomes because of their extended repetitive structure. These two factors could explain most or all of the 'missing' sequence length.

Apart from these small differences, we found the chromosomes to be highly consistent with existing cytogenetic data and genetic maps, and largely collinear with the available complete genome assembly of *B. pendula* (Figure S5), suggesting that they accurately represent the structure of the genome. This validates the 'sequencing-only' strategy used here, which was able to reconstruct essentially the entire genome. The only structural errors were the inversion of the chromosome arms in two instances, which was an artefact of the proximity ligation technique. This could be recognized because the chromosome arms were joined by telomeric repeats instead of centromeric repeats; after manually correcting the error, the chromosomes were consistent with the genetic maps (Appendix S1). This approach could therefore be valuable for sequencing other species for which high-density genetic or physical maps are not yet available. This assembly is an excellent resource for accelerating breeding through molecular marker design and mapping candidate genes for important traits of interest, especially in 'Tombul', the most highly valued Turkish variety.

Further high-quality assemblies from different individuals, such as that currently being constructed for 'Jefferson' (Snelling *et al.*, 2018), will be invaluable for identifying the degree of variation and chromosome rearrangement within the hazelnut population.

Comparison of the *C. avellana* genome with other horticultural crops

With the greatly reduced cost of high-throughput sequencing technologies, the genomes of a number of important nut tree species have been sequenced in recent years, such as *Juglans regia* (walnut; Martínez-García *et al.*, 2016) and *Pistacia vera* (pistachio; Zeng *et al.*, 2019). The closest relative of hazelnut for which a complete genome is available is silver birch (*B. pendula*), for which pseudochromosomes were generated by anchoring genome scaffolds on to a high-density genetic map (Salojärvi *et al.*, 2017); a genetic map and synteny analysis were also used to assemble the *Quercus robur* (oak) genome (Plomion *et al.*, 2018). Among these examples *C. avellana* has both the smallest genome (birch, 440 Mb; walnut and pistachio, approx. 600 Mb; oak, approx. 725 Mb) and the lowest proportion of repetitive elements (38%; the others range from 50–70% of the whole genome). These features make hazelnut an attractive model for genomic studies. Whereas local gene duplications account for over 20% of the genome, as in *B. pendula* and *Q. robur* there was no evidence of any large segmental duplications since the divergence of the Fagales (Salojärvi *et al.*, 2017; Plomion *et al.*, 2018). Tree species are often highly heterozygous, which means that *de novo* sequencing assemblies are often significantly longer than the expected size (Martínez-García *et al.*, 2016; Plomion *et al.*, 2018; Zeng *et al.*, 2019) through some regions being represented twice. We observed the same effect in our assemblies that relied only on short reads; however, incorporating approx. 9.3× genome coverage of long Nanopore reads reduced the assembly to the expected length, suggesting that the heterozygous regions were resolved by this method (Table 1). This gives us confidence that we can make accurate assessments of numbers of functional elements from the *C. avellana* genome.

The number of genes annotated in each of these genomes is comparable, with the 27 270 reported here being very similar to *B. pendula* (28 153) and a little fewer than walnut and pistachio (approx. 32 000 each). Each of these genomes has different gene families that have undergone lineage-specific expansion, however, which may indicate their functional importance to their species. For example, *J. regia* has an unusually large complement of genes for polyphenol synthesis (Martínez-García *et al.*, 2016). Similarly, we found and annotated 147 gene families that were specifically expanded in *C. avellana* (Table S15). The largest of these families included pattern-recognition receptors and protein kinases involved in defence against pathogens; expansion of the immune receptor repertoire has been proposed to be an adaptation to long lifespan in tree species (Plomion *et al.*, 2018). The most significantly enriched gene functions were related to melatonin biosynthesis (20 genes in the genome) and the breakdown of 2,4,6-trinitrotoluene (TNT)-containing compounds (34 genes), however. Melatonin, more widely known as an animal hormone and antioxidant, has been reported to regulate diverse processes in plants in the last few years (Arnao and Hernández-Ruiz, 2018), and is also gaining popularity as a health supplement. TNT is a toxic soil pollutant resulting from the use of explosives that is highly resistant to biodegradation (Rylott *et al.*, 2014). Therefore, closer examination of these gene families could reveal important new uses for *C. avellana*, in the production of nuts with increased nutritional benefits and for the bioremediation of contaminated soil.

Genomic insights into sources of powdery mildew resistance

The genome sequence enabled us to identify 12 full-length *MLO* genes in *C. avellana*, a gene family that is ubiquitous in higher plants and controls susceptibility to powdery mildew (PM) infection in diverse crop species (Acevedo-Garcia *et al.*, 2014). In the plant genomes studied to date, the *MLO* gene family varies in size from eight (wheat) to 39 (soybean)

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

members, and the family is divided into six to eight clades, depending on the species included in the analysis.

All *MLO* genes known to play a role in mildew susceptibility fall into clade IV (monocots) or clade V (dicots). In *A. thaliana*, a loss-of-function mutant of *AtMLO2* confers partial resistance to mildew infection, whereas the triple *AtMLO2 AtMLO6 AtMLO12* mutant is completely resistant. Similarly, the knock-down of *MdMLO19* in apple reduced PM infection by 75% (Pessina *et al.*, 2016). Therefore, although sequence similarity does not guarantee conservation of function, it is likely that one or more of the hazelnut clade-V *MLO* genes could be involved in susceptibility to *E. corylacearum*. In our phylogenetic comparison (Figure 5a), clade VII was close to clade V, and clade IV was basal to both, suggesting that these three may form a subfamily of *MLO* genes involved in PM susceptibility. In apple, *MdMLO19* is upregulated during PM infection, and so is *MdMLO18*, which falls into clade VII (Pessina *et al.*, 2014). Therefore, *Cav02g28680*, along with *Cav07g19540*, *Cav07g19560*, *Cav07g19590* and *Cav10g09690*, are being prioritized for further functional studies. In particular, it would be extremely valuable to identify natural *MLO* mutations leading to PM resistance in hazelnut germplasm, as has been documented in crops such as *Cucumis sativus* (cucumber) and apple (Berg *et al.*, 2015; Pessina *et al.*, 2017).

The MF of MLO proteins is still unclear, but in their absence mildew infection is blocked at the point of cell wall penetration (Kusch and Panstruga, 2017), suggesting that mildew fungi may need to use them as a receptor to initiate cell entry. Their ubiquitous presence in plant genomes and the fact that all naturally occurring *MLO* mutants are recessive indicates that they perform a necessary function for the plant in the absence of PM. This is consistent with our observations of the genomic context and interspecies comparisons of hazelnut *MLO* genes (Figure 5b), where the clade-V and -VII genes in particular seem have been selected for diversification; we hypothesize that historic PM disease pressure could have led to the suppression or disruption of some *MLO* genes, whereas their value in the absence of disease has selected for the duplication and maintenance of new gene copies. With this in mind, further study of the interaction of PM disease and the *MLO* genes in *C. avellana* would provide valuable insight into the function of this gene family in tree species.

A catalogue of allergens suggests strategies for addressing hazelnut sensitization

In the Western world, food allergy is a widely recognized health problem and nut allergy is one of the best-studied examples, with 1–2% of the population having some kind of sensitization to hazelnut (Costa *et al.*, 2015). A diverse group of allergenic proteins have been identified by their ability to provoke an IgE-mediated immune response in sensitized individuals (Figure 6; Table 2). We were able to identify a complete catalogue of genes for these proteins within the *C. avellana* genome; previously reported allergen isoforms result both from multiple genes within the genome and multiple alleles of those genes in different individuals. Based on this catalogue, we also identified likely cross-reactive allergens that have not been reported previously, which will help to guide continuing studies aiming to treat or prevent hazelnut allergy.

The allergic response is complex and varies between individuals. For example, the allergens Cor a 1, Cor a 2, Cor a 6 and Cor a 10 are most abundant in pollen, and it is thought that sensitization to these proteins primarily occurs at the mucosal membranes of the respiratory system, leading to localized symptoms. Sensitization to consumed nuts is more likely to result in severe, systemic allergic responses, however; Cor a 8 and the seed storage proteins (Cor a 9, Cor a 11 and Cor a 14) have each been associated with a higher risk of severe allergy, depending on the study population (Schocker *et al.*, 2004; Garino *et al.*, 2010; Datema *et al.*, 2015). Moreover, many individuals show sensitization to both pollen and nuts. This complex response makes it difficult to predict which proteins could be allergenic; however, some common features can be noted between the known hazelnut allergens (Figure 6b). Cor a 1, Cor a 8 and Cor a TLP are all members of 'pathogenesis response' protein families (PR-10, PR-14 and PR-5, respectively), which were first identified by their increased expression during the plant hypersensitive response to infection. They have diverse MFs but are relatively small proteins that resist protease degradation and are often stabilized by disulfide bonds, meaning that they are likely to be presented to the immune system with their three-dimensional structure intact. Cor a 1, the major hazelnut pollen allergen, includes multiple isoallergens that cross-react with each other and with those from other species (Lüttkopf *et al.*, 2002). From the *C. avellana* genome sequence we were able to identify several new Cor a 1 isoforms and isoallergens, as well as demonstrating that most of the Cor a 1 allergen family members underwent multiplication in the last 5–10 Myr, specifically in the *Corylus* lineage (Figure 6a). Surprisingly, Cor a 1.04 was indicated to have diverged from a birch Bet v 1 homologue (Bet0820g14) 22–43 Mya, significantly after the separation of the *Betula* and *Corylus* genera (Figure 3a). This may indicate a horizontal gene transfer between these species, made possible by the functional similarity between the proteins and their presence in pollen. Further investigation with a population of individuals from each species would be valuable to test this hypothesis.

The expression of multiple, closely related genes in pollen or nuts could increase the risk of allergenicity. Notably, most of the Cor a 1 genes were expressed in our RNA-seq data (Table S13), whereas the more divergent Cor a 1-like genes were not. The TLP family also had multiple

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

expressed genes in *C. avellana*; although not conclusively proven to be allergens in hazelnut (Palacín *et al.*, 2012), they should be regarded as high risk.

The existence of so many variants in the genome suggests that removing these allergens through breeding or genome editing would be impractical. Cor a 8 is a more promising target, however. Although the nsLTPs are also a multigene family, Cor a 8 is highly diverged from the other members, so cross-reactivity is unlikely. Even so, more functional studies are needed to determine whether it is essential to the health of the tree.

The remaining nut allergens are both resistant to the heat used in cooking and the acidity and protease activity of the gastrointestinal tract. The seed storage proteins are highly abundant, making up >50% of all protein in nuts (Beyer *et al.*, 2002). They are also found in condensed 'protein bodies' within the plant cell, increasing the probability that at least some of these proteins will be presented to the immune system with their conformational epitopes intact. Similarly, the oleosins are tightly associated with intracellular oil droplets, which may help to protect them from degradation (Akkerdaas *et al.*, 2006). They were only recognized as allergens relatively recently because the methods used to produce protein extracts from nuts often eliminate oil droplets; however, Cor a 12 sensitivity was observed consistently in 10–25% of hazelnut allergic patients across Europe (Datema *et al.*, 2015).

This illustrates that although the empirical testing of the allergic response is essential to characterize allergens, there is always a risk of overlooking some important factors. In contrast, the genomic survey presented here can give confidence that all relevant proteins have been identified and provide a foundation for further studies of allergenicity.

CONCLUSIONS

We present here a chromosome-level reference genome assembly and annotation for the European hazelnut, *C. avellana* cv. Tombul. Using a combination of short-read, long-read and proximity ligation sequencing we produced a genome of similar quality to those obtained by anchoring contigs to high-density genetic maps; to our knowledge, making this the most complete tree nut genome published to date. The genes and functional elements identified here provide a foundation for ensuring the sustainability of future hazelnut production, for example by identifying targets for breeding or gene knockout that could confer resistance to PM disease and decrease the risk of hazelnut allergy.

EXPERIMENTAL PROCEDURES

Plant materials and nucleic acid purification

Two-year old saplings of *C. avellana* L. var. Tombul were obtained from commercial nurseries in Turkey and cultivated on the Sabanci University campus. The isolation of high-quality genomic DNA (gDNA) proved difficult because of the abundance of polysaccharides and other compounds in hazel tissues that are not easily separated from DNA by standard techniques. Therefore, we adopted an isolation method previously developed for *Betula nana* (Wang *et al.*, 2013), with some modifications. The same protocol was subsequently adapted for RNA extraction (Appendix S1). The best results were obtained by isolating DNA and RNA from fresh leaf buds. The concentration and purity of isolated nucleic acids were estimated by UV spectrophotometry at wavelengths of 260 and 280 nm using the NanoDrop ND-2000c (ThermoFisher Scientific, https://www.thermofisher.com) and checked for degradation by agarose gel electrophoresis. Final DNA concentration was measured using a double-stranded DNA (dsDNA)-specific fluorescent dye (Quant-iT HS dsDNA Assay Kit; ThermoFisher Scientific).

Next-generation sequencing and *de novo* genome sequence assembly

Illumina library preparation and sequencing were carried out by Macrogen (https://www.macrogen.com). For genome sequencing, two paired-end shotgun sequencing libraries were produced using TruSeq Library Preparation kits, size selected to have an average insert of 700–800 bp, and sequenced on a HiSeq 4000 instrument (Illumina, https://www.illumina.com). Similarly, two paired-end transcriptome libraries were prepared using RNA isolated from leaf buds under normal growth conditions, using the TruSeq RNA Sample Prep Kit v2 (Appendix S1; Table S6). Read lengths were 150 nt for genome libraries and 100 nt for transcriptome libraries.

Single-molecule sequencing was carried out in-house; whole genomic DNA was physically disrupted into fragments of approx. 8 kb using a Covaris g-TUBE (Covaris, https://www.covaris.com) and then prepared for NanoPore sequencing on the MinION platform using the Ligation Sequencing Kit 1D, according to the manufacturer's protocols (Oxford NanoPore Technologies, https://nanoporetech.com). Data were obtained from a 48-h run on a single R9.4 flow cell. Proximity ligation sequencing was carried out by Dovetail Genomics (https://dovetailgenomics.com) using their proprietary Chicago™ and HiC™ protocols.

Sequence-read data processing, *de novo* sequence assembly and transcriptome mapping were carried out on the High-Performance Computing cluster (HPC) at Sabanci University, as described in Appendix S1.

Assessment of assembly completeness and genome sequence comparisons

The completeness and accuracy of genome assemblies was assessed using BUSCO 3 (Waterhouse *et al.*, 2018) using default settings, with the reference database for single-copy genes found in land plants (embryophyta) from OrthoDB 9.1 (Kriventseva *et al.*, 2015). The draft *C. avellana* cv. 'Jefferson' genome, coding sequence (CDS) and annotations were retrieved from the original producers' web portal (http://hazelnut.data.mocklerlab.org). The tree genome assemblies used for inter-species comparison were as follows (with GenBank assembly IDs): *Betula nana* (GCA_000327005.1), *Betula pendula* (GCA_900184695.1), *Juglans nigra* (GCA_003123865.1), *Juglans regia* (GCF_001411555.1), *Malus domestica* (GCF_002114115.1), *Populus trichocarpa* (GCA_000002775.3), *Prunus persica* (GCA_000346465.2) and *Quercus robur* (GCA_900291515.1).

Routine sequence similarity searches for single or moderate numbers of sequence elements in the genome assemblies were carried out using standalone BLAST+ 2.2.30 (Camacho *et al.*, 2009).

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

Whole-genome searches and comparisons were realized with BWA 0.7.12 (Li and Durbin, 2010) and resulting alignment files were processed and analysed using SAMTOOLS 1.8 (Li *et al.*, 2009) and BCFTOOLS 1.8 (Danecek *et al.*, 2011).

Annotation of non-coding elements in the *C. avellana* genome

Functional elements in the chromosome pseudomolecules were annotated using a combination of sequence similarity and predictive methods, described in detail in Appendix S1. Briefly, repetitive elements were identified using REPEATMASKER 4.0.7 and a custom repeat database combining novel repeats detected in the Tombul genome with eudicot repetitive elements from RepBase update 22.08 (5913 elements) and mipsREdat 9.3 (26 123 elements) (Smit *et al.*, n.d.; Jurka *et al.*, 2005; Nussbaumer *et al.*, 2012).

Genes encoding tRNAs were identified using TRNASCAN-SE 2.0.0 (Chan and Lowe, 2019). Ribosomal RNA genes were detected by BLASTN searches on the masked chromosome sequences, whereas putative conserved miRNA genes were predicted using all plant miRNAs from miRBase 22 (Kozomara *et al.*, 2019) and the SUMIRFIND/SUMIRFOLD pipeline (Lucas and Budak, 2012). For inter-species comparison, the same methods were used to identify miRNA genes in other plant genomes. Putative miRNA targets were identified in unspliced transcript sequences for high-confidence gene models using PSRNATARGET (Dai *et al.*, 2018).

The GO annotations for all putative miRNA targets (see below) were simplified to the GOslim-plant list. Single enrichment analysis (SEA) of GO terms was carried out on the AgriGO 2.0 web server (Tian *et al.*, 2017); statistical significance was assessed using Fisher's exact test, with terms showing >10% enrichment at an FDR-adjusted P value of <0.05 marked as enriched.

Protein-coding gene modelling and annotation

AUGUSTUS was used to generate *ab initio* gene predictions from the masked Tombul genome based on hidden Markov models optimized for *A. thaliana* (Stanke and Waack, 2003). These predictions were tested by aligning the gene models with the Tombul transcriptome assembly to provide evidence for gene expression; however, because our RNA-seq libraries were only generated from green buds under normal growth conditions, many tissue- and stress-specific genes would not have been expressed. Therefore, the gene models were also assigned to the Viridiplantae OGs conserved and recorded in OrthoMCL-DB (Chen *et al.*, 2006). Low-quality matches were eliminated from the annotation and gene models in some OGs: e.g., those encoding retroviral proteins were manually curated (Appendix S1; Table S2). Where the intron–exon structure of *C. avellana* gene models was inconsistent with orthologues from other plants, fresh predictions were generated using FGENESH (Solovyev *et al.*, 2006), with training parameters from *B. nana* and the most consistent gene model retained. The automated annotation of the gene models with InterPro protein domains, and ontological terms from MAPMAN and GO, was carried out using MERCATOR 4 and TRAPID 2.0 (Van Bel *et al.*, 2013; Schwacke *et al.*, 2019). In addition, BLASTP searches of predicted protein sequences against the National Center for Biotechnology Information (NCBI) non-redundant protein database were used to find the closest orthologue for each gene model.

Inter-species gene comparison and phylogenetic analysis

Complete CDS and protein sequences for all predicted genes of silver birch (*B. pendula*), English walnut (*J. regia* cv. 'Chandler' assembly v1.0), pedunculate oak (*Q. robur* pseudohaploid PM1N), apple (*M. domestica* GDDH13 v1.1) and peach (*Prunus persica* vJGI 1.0) were retrieved from www.hardwoodgenomics.org and www.rosaceae.org (Jung *et al.*, 2019; HWG, 2019). The complete predicted proteomes of *C. avellana* var. Tombul and these five species were clustered into OGs using ORTHOFINDER 2.3.11 (Emms and Kelly, 2019). Using the all-against-all protein BLAST output from ORTHOFINDER, regions of collinearity/synteny within and between genomes were identified with MCSCANX (Wang *et al.*, 2012). Each pair of matching sequences was aligned using CLUSTALW (Thompson *et al.*, 1994) and pairwise substitution rates (K_s and K_a) calculated using a custom PERL script. For Bayesian estimation of divergence times, the MCMCTREE algorithm from PAML 4.9j (Yang, 2007) was applied with fossil calibrations on four nodes, described in detail in Appendix S1 (Table S3). Time trees were drawn using FIGTREE 1.4.2 (Rambaut, 2018).

Characterization of *MLO* and allergen gene families

All gene models annotated with the InterPro domains IPR004326, 'MIo-related protein', IPR000916, 'Bet v I allergen' and other allergen-related domains were inspected individually. *Arabidopsis thaliana* MLO protein sequences were obtained from the Araport11 genome annotation (Cheng *et al.*, 2017), whereas those for *M. domestica* came from the Genome Database for Rosaceae (Jung *et al.*, 2019). Reference allergen sequences were retrieved from GenBank, using the accessions listed at www.allergen.org. *MLO* and allergen gene models were verified and re-annotated as described in Appendix S1. The transmembrane structure of predicted MLO protein sequences was evaluated using TMHMM 2.0 with default parameters (Krogh *et al.*, 2001). Incomplete gene fragments and unexpressed homologues were detected by searching the pseudochromosomes using TBLASTN, with the complete protein sequences as the query. MEGA 6.0 (Tamura *et al.*, 2013) was used for multiple sequence alignments (using the MUSCLE algorithm) and for phylogenetic clustering.

DATA AVAILABILITY STATEMENT

All raw and processed sequencing data generated in this study have been submitted to the European Nucleotide Archive (ENA), under BioProject accession number PRJEB31933 (https://www.ebi.ac.uk/ena/data/view/PRJEB31933). The final chromosome assembly accession is GCA_901000735. Gene model transcript, CDS and amino acid sequences, and annotations for other functional elements are available through www.hardwoodgenomics.org. All other relevant data are included within the article and supporting information.

ACKNOWLEDGMENTS

This work was supported financially by the Scientific and Technological Research Council of Turkey (TÜBITAK grant no. 215O446 to SJL) and by the Newton Fund Institutional Links programme (grant no. 216394498 to RJAB). This study used the Sabanci HPC Cluster for computing support. The authors would like to thank Nihal Öztolan Erol and Andrew J. Helmstetter for helpful discussions.

AUTHOR CONTRIBUTIONS

SJL conceived the study, developed analysis pipelines, analysed and interpreted the data, and wrote the article. KK and BA acquired, analysed and interpreted the data, prepared the figures and drafted sections of the article. RJAB contributed substantively to the study design and revised the article. IB analysed and interpreted additional

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

data. All authors read and approved the final version for publication.

CONFLICT OF INTEREST

The authors declare that they have no competing interests associated with this work.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Detailed description of data analysis pipelines, annotation and comparative genomics analyses.

Figure S1. Comparison of SSR marker order on chromosomes with genetic maps.

Figure S2. Relative abundance of interspersed repeat families in the *Corylus avellana* genome.

Figure S3. Relative frequency of tRNA anticodons in *Corylus avellana* var. 'Tombul' and other plant genomes.

Figure S4. Genome statistics of non-coding RNA elements in *Corylus avellana*.

Figure S5. Dot plot representing whole-genome alignment of the *Corylus avellana* and *Betula pendula* chromosome assemblies.

Figure S6. Multiple sequence alignment of predicted MLO protein sequences.

Figure S7. Phylogenetic inference (ML method) of tree *MLO*-like genes.

Table S1. PFAM protein domains used for LTR identification.

Table S2. Orthologue groups containing transposable element domains in the *Corylus avellana* gene set.

Table S3. Fossil calibrations used for the calculation of time trees.

Table S4. Genome sequencing and assembly statistics.

Table S5. Chromosome pseudomolecule statistics.

Table S6. Sequencing and assembly statistics for RNA-seq libraries.

Table S7. Summary of repetitive elements found in the *Corylus avellana* genome.

Table S8. Identity and locations of predicted tRNA genes in the *Corylus avellana* genome.

Table S9. Comparison of tRNA gene codon frequencies.

Table S10. Best BLAST matches to *45S* and *5S* rRNA genes and ribosomal proteins on *Corylus avellana* pseudochromosomes.

Table S11. Predicted miRNA genes, locations and pre-miRNA characteristics in *Corylus avellana* genome.

Table S12. Target prediction for hazelnut miRNAs in high-confidence protein-coding gene models.

Table S13. Summary of all high-confidence protein models annotated in the *Corylus avellana* genome, their location, conserved domains and best match to known plant proteins.

Table S14. Assignment of *Corylus avellana* predicted proteins to MAPMAN functional bins.

Table S15. Significantly enriched GO terms in gene families expanded in the *Corylus* lineage.

Table S16. MLO-related gene model predictions and results of re-annotation.

Table S17. Predicted genes encoding known and newly identified hazelnut allergen proteins.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://www.ebi.ac.uk/ena/browser/view/PRJEB31933.

REFERENCES

Acevedo-Garcia, J., Kusch, S. and Panstruga, R. (2014) Magical mystery tour: MLO proteins in plant immunity and beyond. *New Phytol.* **204**, 273–281.

Akkerdaas, J.H., Schocker, F., Vieths, S. *et al.* (2006) Cloning of oleosin, a putative new hazelnut allergen, using a hazelnut cDNA library. *Mol. Nutr. Food Res.* **50**, 18–23.

Aquilano, K., Ceci, V., Gismondi, A. *et al.* (2019) Adipocyte metabolism is improved by TNF receptor-targeting small RNAs identi fi ed from dried. *Commun. Biol.* **2**, 317.

Arnao, M.B. and Hernández-Ruiz, J. (2018) Melatonin and its relationship to plant hormones. *Ann. Bot.* **121**, 195–207.

Avsar, B. and Aliabadi, D.E. (2017) Identification of microRNA elements from genomic data of European hazelnut (*Corylus avellana* L.) and its close relatives. *Plant Omi. J.* **10**, 190–196.

Bel, M.V., Proost, S., Neste, C.V., Deforce, D., Van de Peer, Y.V. and Vandepoele, K. (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. *Genome Biol.* **14**, R134.

Berg, J.A., Appiano, M., Santillán Martínez, M., Hermans, F.W.K., Vriezen, W.H., Visser, R.G.F., Bai, Y. and Schouten, H.J. (2015) A transposable element insertion in the susceptibility gene CsaMLO8 results in hypocotyl resistance to powdery mildew in cucumber. *BMC Plant Biol.* **15**, 243.

Beyer, K., Grishina, G., Bardina, L., Grishin, A. and Sampson, H.A. (2002) Identification of an 11S globulin as a major hazelnut food allergen in hazelnut-induced systemic reactions. *J. Allergy Clin. Immunol.* **110**, 517–523.

Bhattarai, G. and Mehlenbacher, S.A. (2017) In silico development & characterization of tri-nucleotide simple sequence repeat markers in hazelnut (*Corylus avellana* L.). *PLoS One*, **12**, e0178061.

Breiteneder, H., Ferreira, F., Hoffmann-Sommergruber, K., Ebner, C., Breitenbach, M., Rumpold, H., Kraft, D. and Scheiner, O. (1993) Four recombinant isoforms of Cor a I, the major allergen of hazel pollen, show different IgE-binding properties. *Eur. J. Biochem.* **212**, 355–362.

Büschges, R., Hollricher, K., Panstruga, R. *et al.* (1997) The barley Mlo gene: a novel control element of plant pathogen resistance. *Cell*, **88**, 695–705.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chan, P.P. and Lowe, T.M. (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods in Molecular Biology* **1962**, 1–14. https://doi.org/10.1007/978-1-4939-9173-0_1

Chapman, M.D., Pomés, A., Breiteneder, H. and Ferreira, F. (2007) Nomenclature and structural biology of allergens. *J. Allergy Clin. Immunol.* **119**, 414–420.

Chen, F., Mackey, A.J., Stoeckert, C.J. and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363–D368.

Chen, X., Zhang, J., Liu, Q., Guo, W., Zhao, T., Ma, Q. and Wang, G. (2014) Transcriptome sequencing and identification of cold tolerance genes in hardy Corylus species (C. heterophylla fisch) floral buds. *PLoS One*, **9**, e108604.

Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S. and Town, C.D. (2017) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804.

Colburn, B.C., Mehlenbacher, S.A. and Sathuvalli, V.R. (2017) Development and mapping of microsatellite markers from transcriptome sequences of European hazelnut (*Corylus avellana* L.) and use for germplasm characterization. *Mol. Breed.* **37**, 16.

Costa, J., Mafra, I., Carrapatoso, I. and Oliveira, M.B.P.P. (2015) Hazelnut allergens: molecular characterization, detection, and clinical relevance. *Crit. Rev. Food Sci. Nutr.* **56**, 2579–2605.

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

Dai, X., Zhuang, Z. and Zhao, P.X. (2018) psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* **46**, W49–W54.

Danecek, P., Auton, A., Abecasis, G. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Datema, M.R., Zuidmeer-Jongejan, L., Asero, R. *et al.* (2015) Hazelnut allergy across Europe dissected molecularly: a EuroPrevall outpatient clinic survey. *J. Allergy Clin. Immunol.* **136**, 382–391.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238.

Falistocco, E. and Marconi, G. (2013) Cytogenetic characterization by in situ hybridization techniques and molecular analysis of 5S rRNA genes of the European hazelnut (*Corylus avellana*). *Genome*, **56**, 155–159.

FAO (2017) *FAOSTAT*. Available at: http://www.fao.org/faostat/en/#data/QC Accessed August 11, 2017.

Fischer, S., Brunk, B.P., Chen, F. *et al.* (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinforma. Chapter*, **6**, Unit 6.12.1-19.

Garino, C., Zuidmeer, L., Marsh, J. *et al.* (2010) Isolation, cloning, and characterization of the 2S albumin: a new allergen from hazelnut. *Mol. Nutr. Food Res.* **54**, 1257–1265.

Gürcan, K. and Mehlenbacher, S.A. (2010) Development of microsatellite marker loci for European hazelnut (*Corylus avellana* L.) from ISSR fragments. *Mol. Breed.* **26**, 551–559.

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

Hoffmann-Sommergruber, K., Vanek-Krebitz, M., Radauer, C., Wen, J., Ferreira, F., Scheiner, O. and Breiteneder, H. (1997) Genomic characterization of members of the Bet v 1 family: genes coding for allergens and pathogenesis-related proteins share intron positions. *Gene*, **197**, 91–100.

HWG. (2019) No Title. Hardwood Genomics Proj. Available at: https://www.hardwoodgenomics.org/ (accessed September 11, 2020).

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S. *et al.* (2012) A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3.

Jimenez-Lopez, J.C., Morales, S., Castro, A.J., Volkmann, D., Rodríguez-García, M.I. and de Alché, J.D. (2012) Characterization of profilin polymorphism in pollen with a focus on multifunctionality. *PLoS One*, **7**, e30878.

Jung, S., Lee, T., Cheng, C.H. *et al.* (2019) 15 years of GDR: new data and functionality in the genome database for Rosaceae. *Nucleic Acids Res.* **47**, D1137–D1145.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Kozomara, A., Birgaoanu, M. and Griffiths-Jones, S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162.

Kriventseva, E.V., Tegenfeldt, F., Petty, T.J., Waterhouse, R.M., Simão, F.A., Pozdnyakov, I.A., Ioannidis, P. and Zdobnov, E.M. (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.* **43**, D250–D256.

Krogh, A., Larsson, B., Heijne, G.V. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

Kusch, S. and Panstruga, R. (2017) Mlo-based resistance: An apparently universal "weapon" to defeat powdery mildew disease. *Mol. Plant-Microbe Interact.* **30**, 179–189.

Larson-Johnson, K. (2016) Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. *New Phytol.* **209**, 418–435.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lucas, S.J. and Budak, H. (2012) Sorting the wheat from the Chaff: identifying miRNAs in genomic survey sequences of *Triticum aestivum* chromosome 1AL. *PLoS One*, **7**, e40859.

Lüttkopf, D., Müller, U., Skov, P.S. *et al.* (2002) Comparison of four variants of a major allergen in hazelnut (*Corylus avellana*) Cor a 1.04 with the major hazel pollen allergen Cor a 1.01. *Mol. Immunol.* **38**, 515–525.

Ma, H., Lu, Z., Liu, B., Qiu, Q. and Liu, J. (2013) Transcriptome analyses of a Chinese hazelnut species *Corylus mandshurica*. *BMC Plant Biol.* **13**, 152.

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L.L. and Hernández-Hernández, T. (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 407–453.

Marinoni, D.T., Beltramo, C., Akkak, A., Destefanis, M.L., Boccacci, P. and Botta, R. (2009) Gene expression and sporophytic self-incompatibility in hazelnut. *Acta Hortic.*, **845**, 227–232.

Marinoni, D.T., Valentini, N., Portis, E., Acquadro, A., Beltramo, C., Mehlenbacher, S.A., Mockler, T.C., Rowley, E.R. and Botta, R. (2018) High density SNP mapping and QTL analysis for time of leaf budburst in *Corylus avellana* L. *PLoS One*, **13**, e0195408.

Martínez-García, P.J., Crepeau, M.W., Puiu, D. *et al.* (2016) The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J.* **87**, 507–532.

Mehlenbacher, S.A., Brown, R.N., Nouhra, E.R., Gökirmak, T., Bassil, N.V. and Kubisiak, T.L. (2006) A genetic linkage map for hazelnut (*Corylus avellana* L.) based on RAPD and SSR markers. *Genome*, **49**, 122–133.

Molnar, T.J. and Capik, J. (2012) Advances in hazelnut research in North America. *Acta Hortic.* **940**, 57–65.

Niu, X.M., Xu, Y.C., Li, Z.W. *et al.* (2019) Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc. Natl. Acad. Sci. USA*, **116**, 6908–6913.

Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H. and Spannagl, M. (2012) MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* **41**, D1144–D1151.

Öztürk, S.C., Göktay, M., Allmer, J., Doğanlar, S. and Frary, A. (2018) Development of simple sequence repeat markers in hazelnut (*Corylus avellana* L.) by next-generation sequencing and discrimination of Turkish hazelnut cultivars. *Plant Mol. Biol. Report.* **36**, 800–811.

Palacín, A., Rivas, L.A., Gómez-Casado, C. *et al.* (2012) The involvement of thaumatin-like proteins in plant food cross-reactivity: a multicenter study using a specific protein microarray. *PLoS One*, **7**, e44088.

Pessina, S., Angeli, D., Martens, S., Visser, R.G.F., Bai, Y., Salamini, F., Velasco, R., Schouten, H.J. and Malnoy, M. (2016) The knock-down of the expression of MdMLO19 reduces susceptibility to powdery mildew (*Podosphaera leucotricha*) in apple (*Malus domestica*). *Plant Biotechnol. J.* **14**, 2033–2044.

Pessina, S., Palmieri, L., Bianco, L. *et al.* (2017) Frequency of a natural truncated allele of MdMLO19 in the germplasm of *Malus domestica*. *Mol. Breed.* **37**, 7.

Pessina, S., Pavan, S., Catalano, D., Gallotta, A., Visser, R.G.F., Bai, Y., Malnoy, M. and Schouten, H.J. (2014) Characterization of the MLO gene family in Rosaceae and gene expression analysis in *Malus domestica*. *BMC Genom.* **15**, 618.

Plomion, C., Aury, J.M., Amselem, J. *et al.* (2018) Oak genome reveals facets of long lifespan. *Nat. Plants*, **4**, 440–452.

Rambaut, A. (2018) FigTree. Mol. Evol. Phylogenetics Epidemiol. Available at: http://tree.bio.ed.ac.uk/software/figtree/ (accessed September 17, 2020).

Reis, M.D., Donoghue, P.C.J. and Yang, Z. (2016) Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**, 71–80.

Rowley, E.R., Fox, S.E., Bryant, D.W., Sullivan, C.M., Priest, H.D., Givan, S.A., Mehlenbacher, S.A. and Mockler, T.C. (2012) Assembly and characterization of the European hazelnut "Jefferson" transcriptome. *Crop Sci.* **52**, 2679–2686.

Rowley, E.R., Vanburen, R., Bryant, D.W., Priest, H.D., Shawn, A. and Mockler, T.C. (2018) A draft genome and high-density genetic map of European Hazelnut (*Corylus avellana* L). bioRxiv 469015; https://doi.org/10.1101/469015

Rylott, E.L., Gunning, V., Tzafestas, K., Sparrow, H., Johnston, E.J., Brentnall, A.S., Potts, J.R. and Bruce, N.C. (2014) Phytodetoxification of the environmental pollutant and explosive 2,4,6-trinitrotoluene. *Plant Signal. Behav.* **10**, e977714.

Salojärvi, J., Smolander, O.P., Nieminen, K. *et al.* (2017) Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat. Genet.* **49**, 904–912.

© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

Sathuvalli, V., Mehlenbacher, S.A. and Smith, D.C. (2017) High-resolution genetic and physical mapping of the eastern filbert blight resistance region in 'Jefferson' Hazelnut (L.). *Plant Genome*, **10**, 1–12.

Schocker, F., Lüttkopf, D., Scheurer, S. *et al.* (2004) Recombinant lipid transfer protein Cor a 8 from hazelnut: a new tool for in vitro diagnosis of potentially severe hazelnut allergy. *J. Allergy Clin. Immunol.* **113**, 141–147.

Schwacke, R., Ponce-Soto, G.Y., Krause, K. *et al.* (2019) MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. Plant*, **12**, 879–892.

Sezer, A., Dolar, F.S., Lucas, S.J., Köse, Ç. and Gümüş, E. (2017) First report of the recently introduced, destructive powdery mildew *Erysiphe corylacearum* on hazelnut in Turkey. *Phytoparasitica*, **45**, 577–581.

Smit, A.F., Hubley, R. and Green, P. (n.d.) RepeatMasker. Available at: http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker (accessed September 4, 2019).

Snelling, J.W., Sathuvalli, V.R., Colburn, B.C., Bhattarai, G., Rowley, E.R., Mockler, T.C., Saski, C.A., Copetti, D. and Mehlenbacher, S.A. (2018) Genomic resource development in hazelnut breeding. *Acta Hortic.* **1226**, 39–46.

Solovyev, V., Kosarev, P., Seledsov, I. and Vorobyev, D. (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, ii215–ii225.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W. and Su, Z. (2017) AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129.

Ustaoğlu, B. (2012) The effect of climatic conditions on hazelnut (*Corylus avellana*) yield in Giresun (TURKEY). *Marmara Geogr. Rev.* **26**, 302–323.

Wang, N., Thomson, M., Bodles, W.J.A., Crawford, R.M.M., Hunt, H.V., Featherstone, A.W., Pellicer, J. and Buggs, R.J.A. (2013) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Mol. Ecol.* **22**, 3098–3111.

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C. and Qiu, J.-L. (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nat. Biotechnol.* **32**, 947–951.

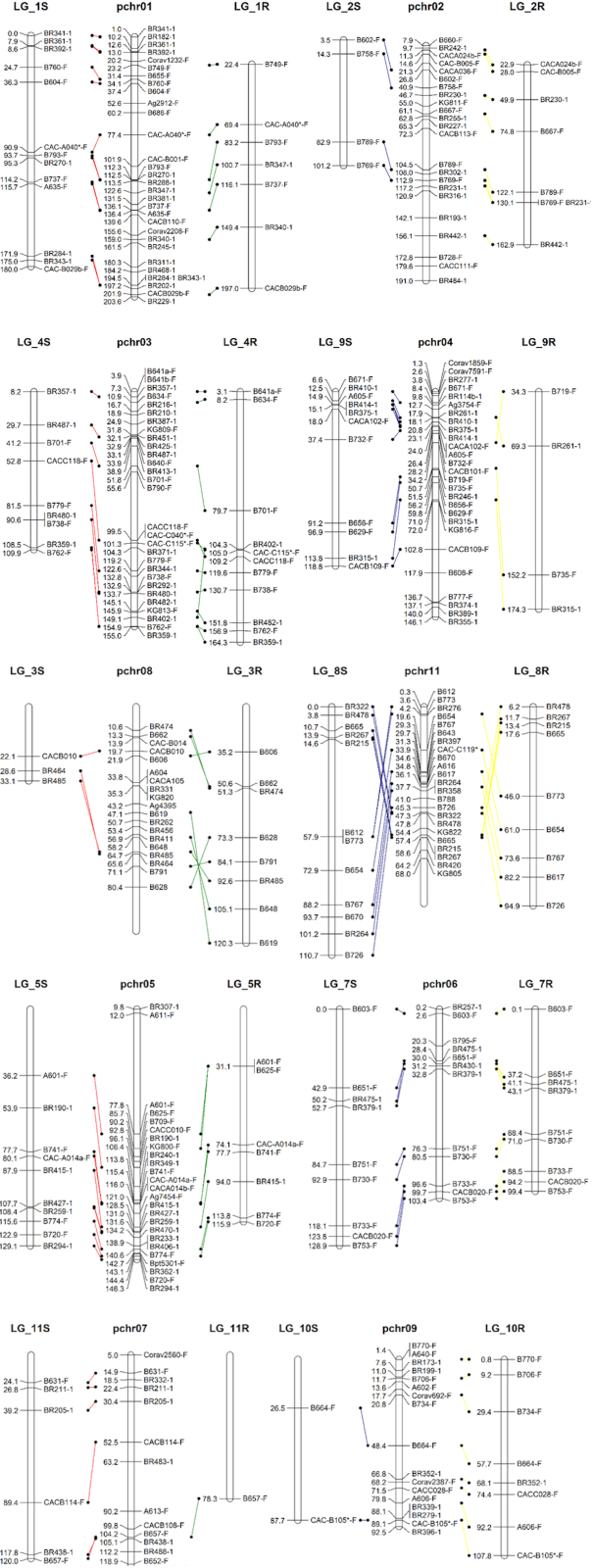
Wang, Y., Tang, H., Debarry, J.D. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

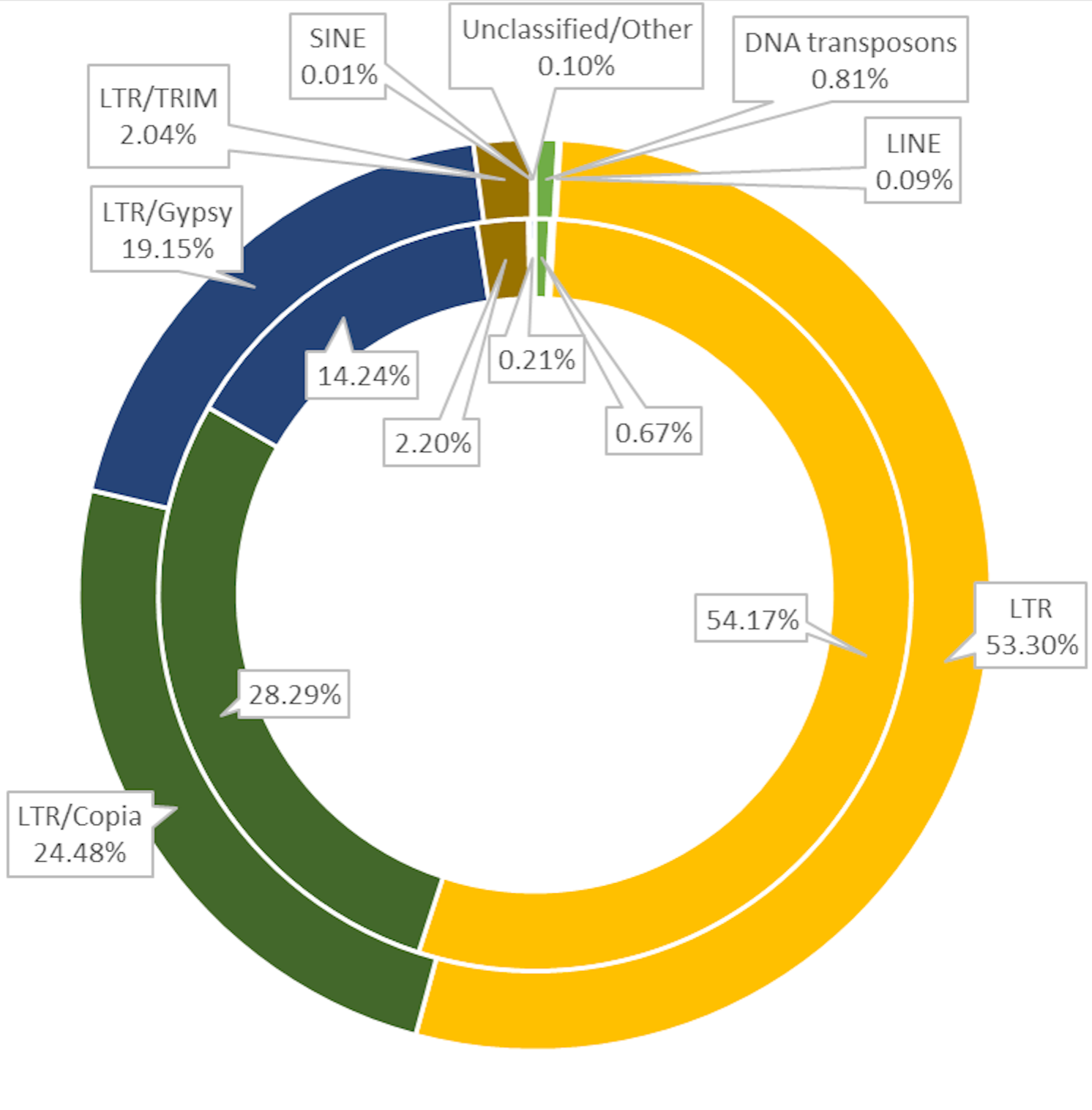
Waterhouse, R.M., Seppey, M., Simao, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. and Zdobnov, E.M. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548.

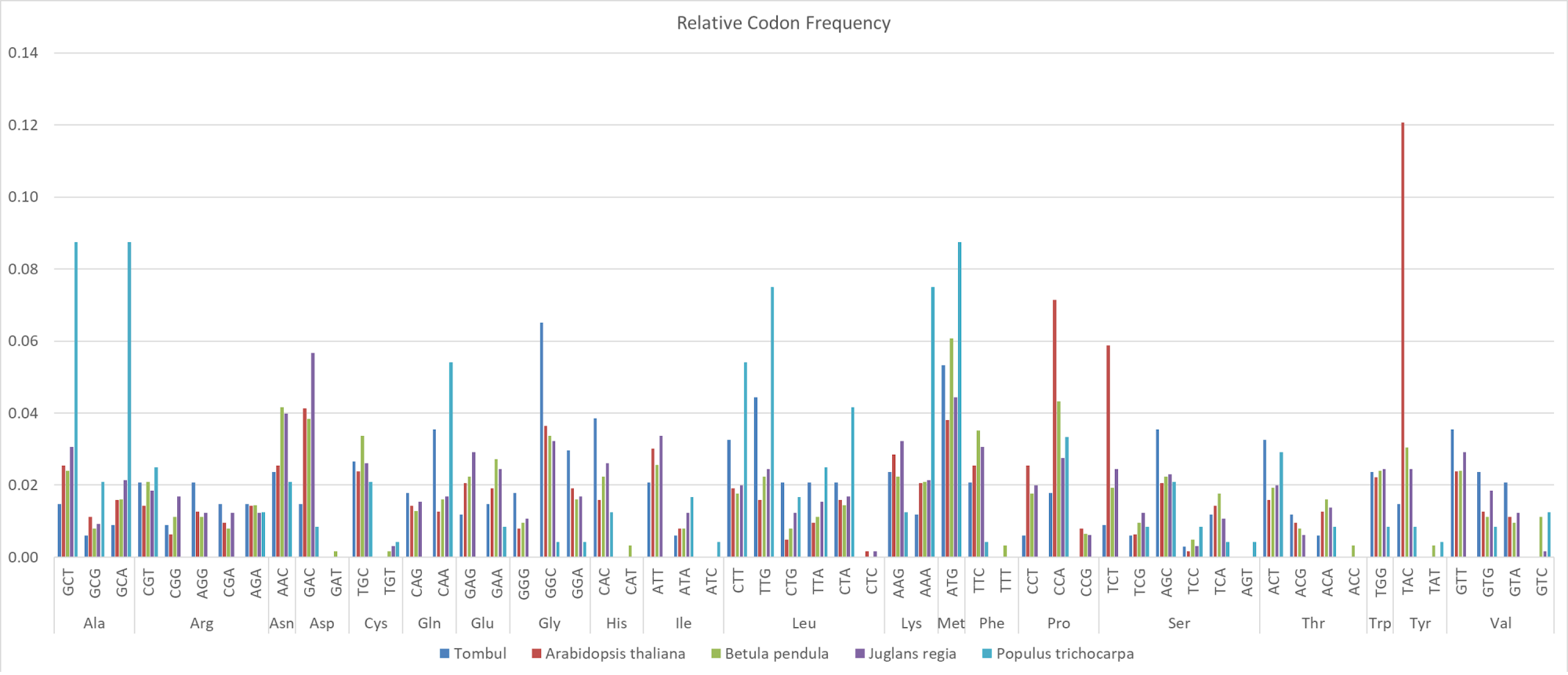
Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

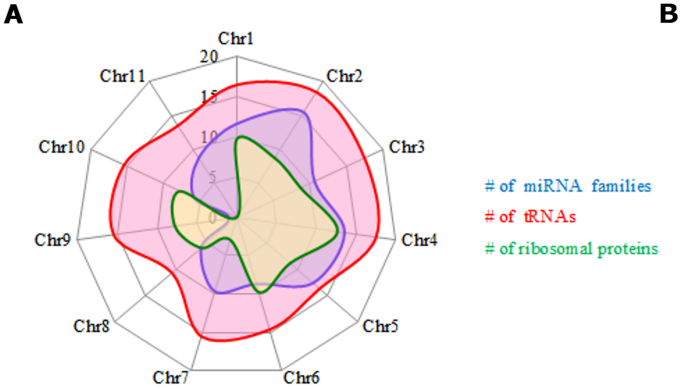
Zeng, L., Tu, X.L., Dai, H. *et al.* (2019) Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* **20**, 79.

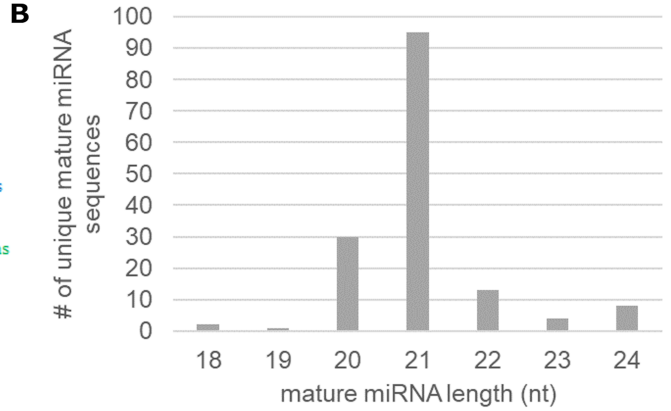
© 2020 Society for Experimental Biology and John Wiley & Sons Ltd,

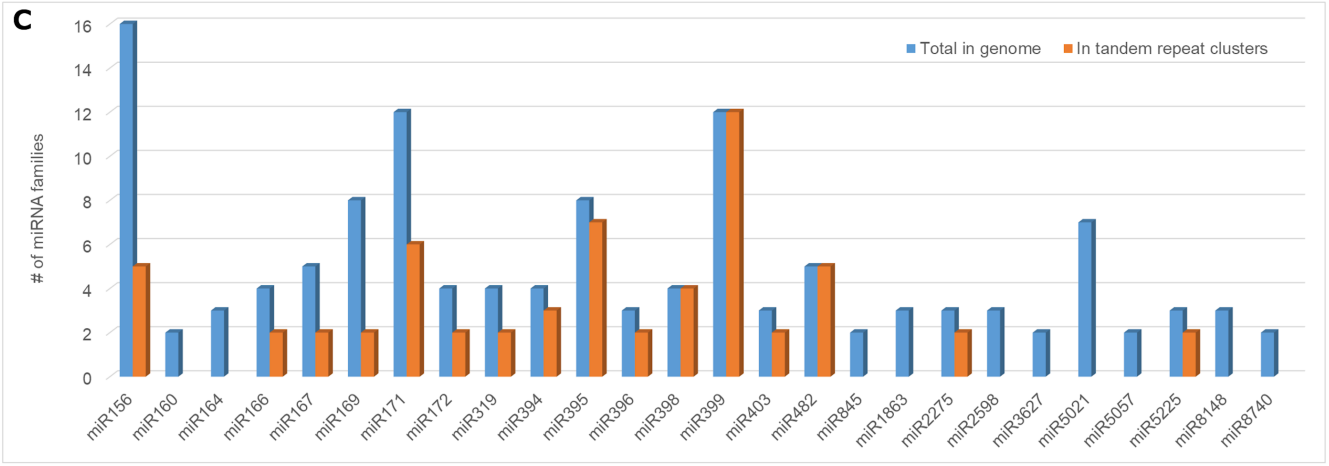

LG_1S
pchr01
LG_1R
LG_2S
pchr02
LG_2R
LG_4S
pchr03
LG_4R
LG_9S
pchr04
LG_9R
LG_3S
pchr08
LG_3R
LG_8S
pchr11
LG_8R
LG_5S
pchr05
LG_5R
LG_7S
pchr06
LG_7R
LG_11S
pchr07
LG_11R
LG_10S
pchr09
LG_10R

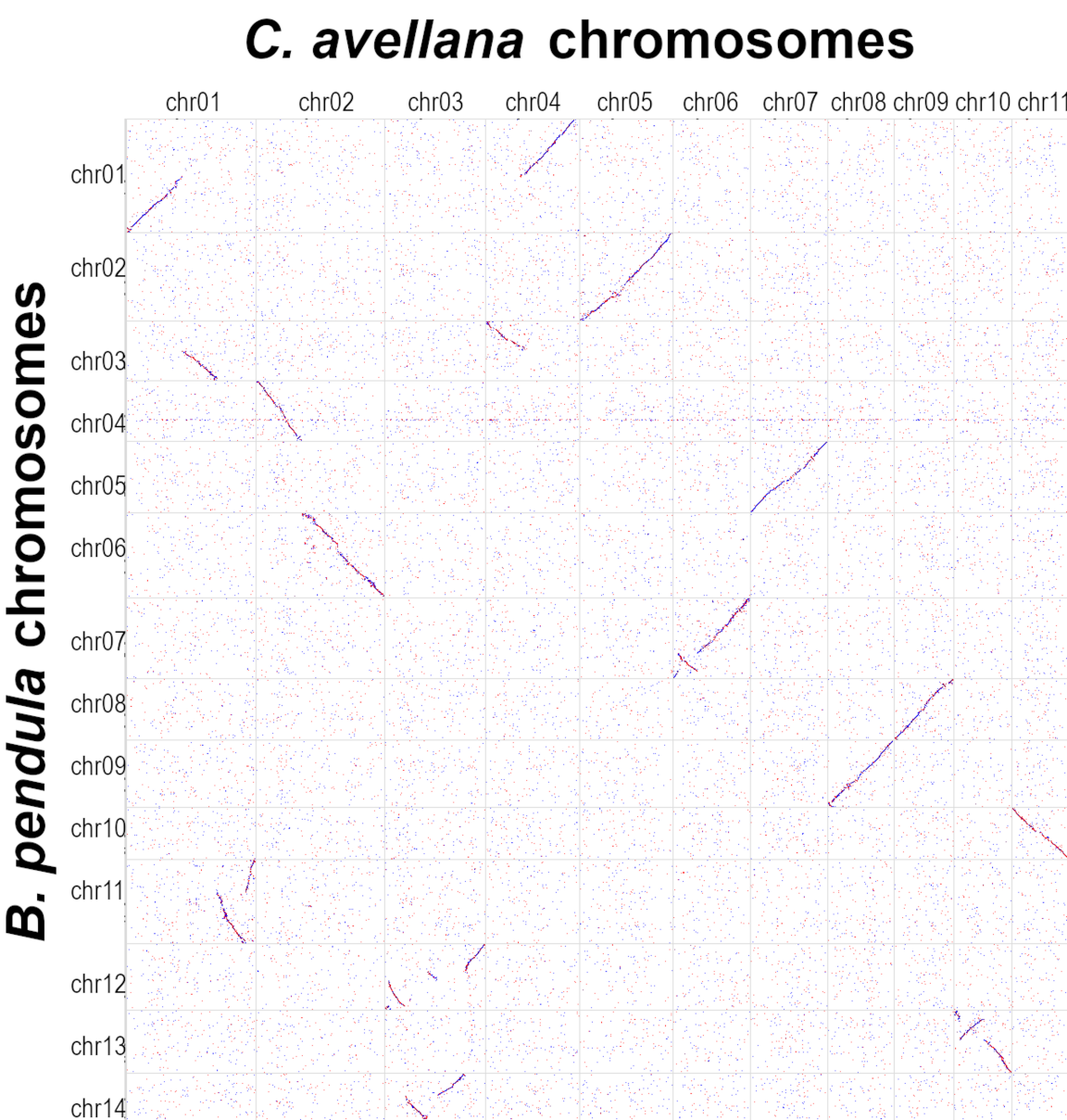

SINE
0.01%
Unclassified/Other
0.10%
DNA transposons
0.81%
LTR/TRIM
2.04%
LINE
0.09%
LTR/Gypsy
19.15%
14.24%
0.21%
2.20%
0.67%
54.17%
LTR
53.30%
28.29%
LTR/Copia
24.48%

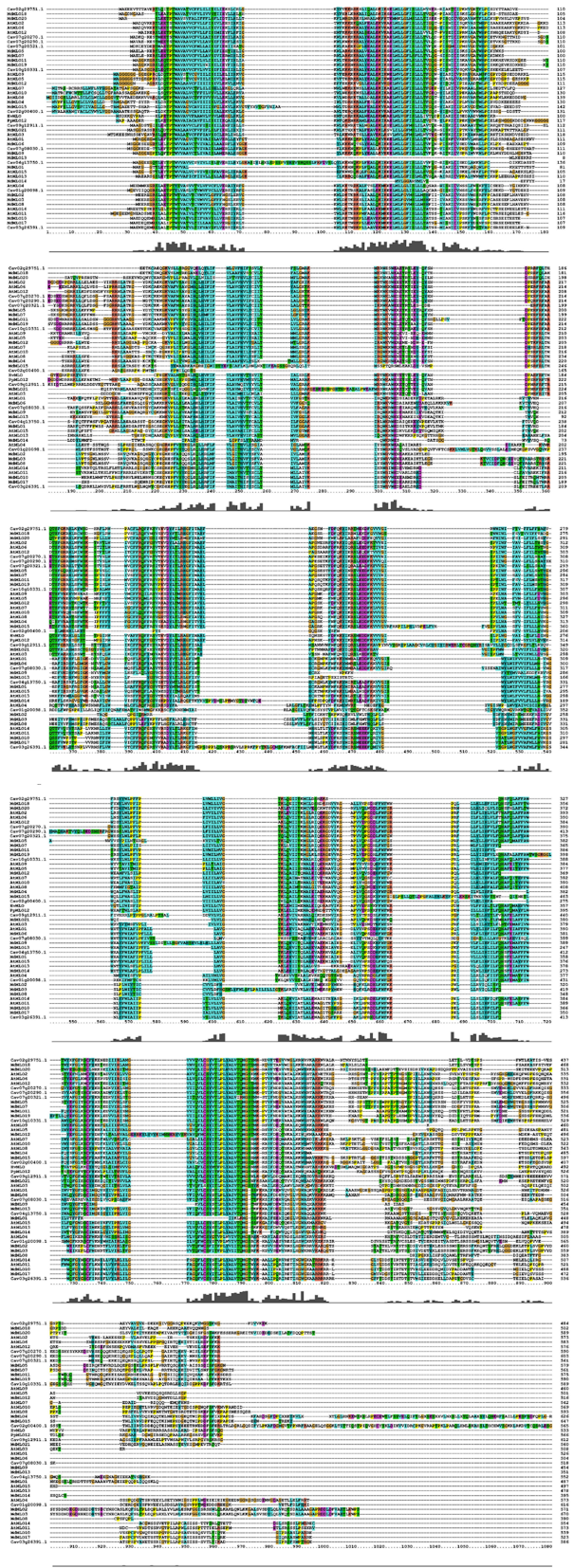

Relative Codon Frequency
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
GCT GCG GCA
CGT CGG AGG CGA AGA
AAC
GAC GAT
TGC TGT
CAG CAA
GAG GAA
GGG GGC GGA
CAC CAT
ATT ATA ATC
CTT TTG CTG TTA CTA CTC
AAG AAA
ATG
TTC TTT
CCT CCA CCG
TCT TCG AGC TCC TCA AGT
ACT ACG ACA ACC
TGG
TAC TAT
GTT GTG GTA GTC
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
Tombul
Arabidopsis thaliana
Betula pendula
Juglans regia
Populus trichocarpa

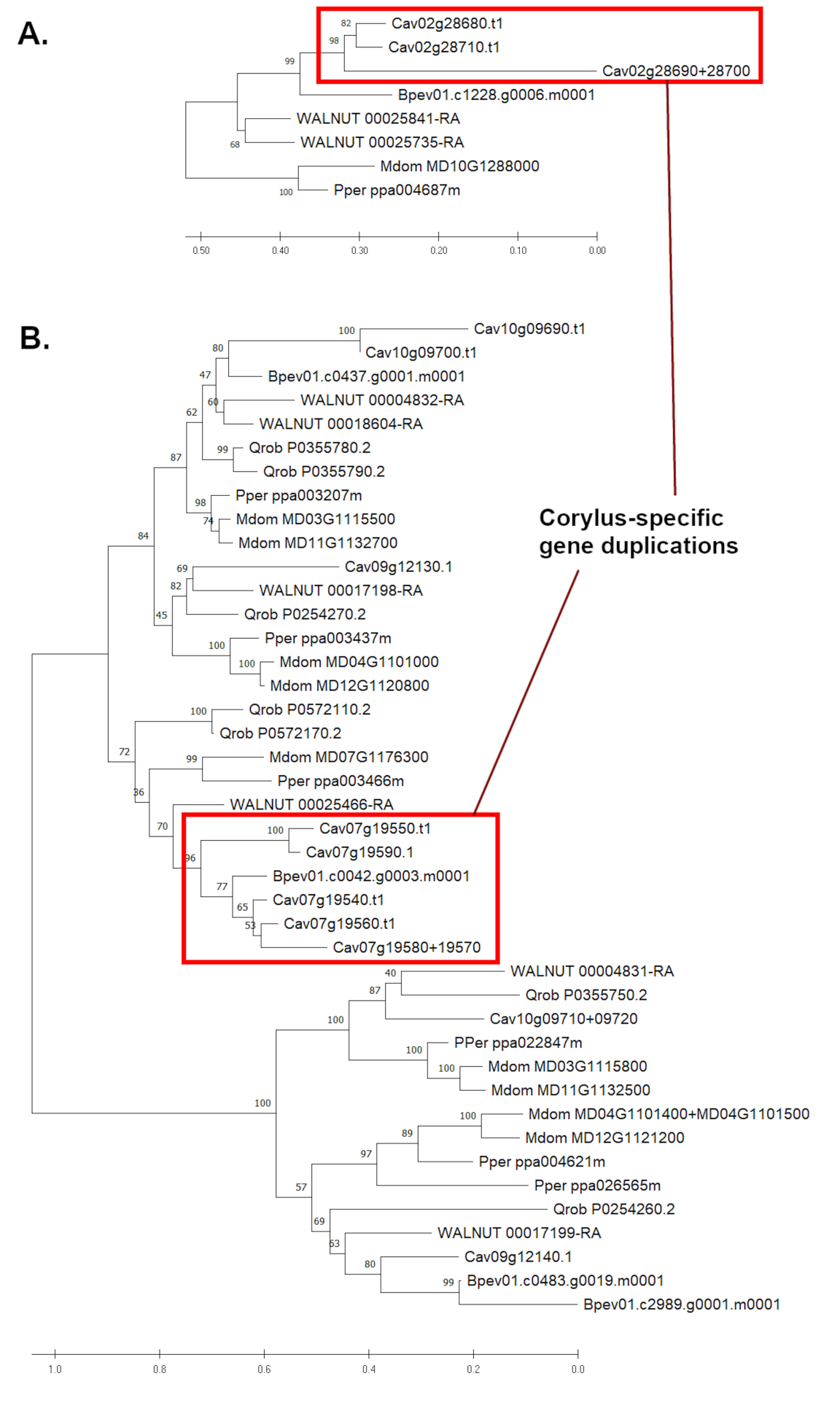

A
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
20
15
10
5
0
of miRNA families
of tRNAs
of ribosomal proteins


B
of unique mature miRNA sequences
100
90
80
70
60
50
40
30
20
10
0
18
19
20
21
22
23
24
mature miRNA length (nt)


C
of miRNA families
16
14
12
10
8
6
4
2
0
Total in genome
In tandem repeat clusters
miR156
miR160
miR164
miR166
miR167
miR169
miR171
miR172
miR319
miR394
miR395
miR396
miR398
miR399
miR403
miR482
miR845
miR1863
miR2275
miR2598
miR3627
miR5021
miR5057
miR5225
miR8148
miR8740


C. avellana chromosomes
B. pendula chromosomes
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14




A.
Cav02g28680.t1
Cav02g28710.t1
Cav02g28690+28700
Bpev01.c1228.g0006.m0001
WALNUT 00025841-RA
WALNUT 00025735-RA
Mdom MD10G1288000
Pper ppa004687m
0.50
0.40
0.30
0.20
0.10
0.00
B.
Corylus-specific gene duplications
Cav10g09690.t1
Cav10g09700.t1
Bpev01.c0437.g0001.m0001
WALNUT 00004832-RA
WALNUT 00018604-RA
Qrob P0355780.2
Qrob P0355790.2
Pper ppa003207m
Mdom MD03G1115500
Mdom MD11G1132700
Cav09g12130.1
WALNUT 00017198-RA
Qrob P0254270.2
Pper ppa003437m
Mdom MD04G1101000
Mdom MD12G1120800
Qrob P0572110.2
Qrob P0572170.2
Mdom MD07G1176300
Pper ppa003466m
WALNUT 00025466-RA
Cav07g19550.t1
Cav07g19590.1
Bpev01.c0042.g0003.m0001
Cav07g19540.t1
Cav07g19560.t1
Cav07g19580+19570
WALNUT 00004831-RA
Qrob P0355750.2
Cav10g09710+09720
PPer ppa022847m
Mdom MD03G1115800
Mdom MD11G1132500
Mdom MD04G1101400+MD04G1101500
Mdom MD12G1121200
Pper ppa004621m
Pper ppa026565m
Qrob P0254260.2
WALNUT 00017199-RA
Cav09g12140.1
Bpev01.c0483.g0019.m0001
Bpev01.c2989.g0001.m0001
1.0
0.8
0.6
0.4
0.2
0.0

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Contents: *Page*

Supplementary Methods

1.1	DNA and RNA purification	3
1.2	Sequence data processing and genome assembly	3
1.3	Chromosome structure verification & map comparison	4
1.4	Telomeric and Centromeric Repeats	5
1.5	Discovery and masking of repetitive elements	5
1.6	Prediction of hazelnut tRNA, rRNA and miRNAs	7
1.7	Construction and annotation of gene models	7
1.8	Data for comparative genomics and phylogenetic tree construction	9
1.9	Ortholog identification and sequence comparisons	9
1.10	Species and gene set divergence time estimation	10

Supplementary Results and Discussion

2.1	Genome sequencing and hybrid assembly approach	12
2.2	Transcriptome Assembly	14
2.3	Chromosome structure verification & map comparison	14
2.4	Detection of Repetitive elements in the *C. avellana* genome	18
2.5	Annotation and inter-species comparison of tRNA genes	20
2.6	Ribosomal RNA and protein distribution	21
2.7	MicroRNA annotation and target identification	21
2.8	Protein-coding genes and annotation	23
2.9	Comparative Genome Alignment	24
2.10	Characterization of the MLO gene family	25
2.11	Identification of Allergen-coding genes and homologs	31

Supplement References 32

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

List of Tables *Page*

Table S1. PFAM Protein domains used for LTR identification. 6

Table S2. Ortholog Groups containing transposable element domains in the *C. avellana* gene set. 9

Table S3. Fossil calibrations used for calculation of timetrees. 11

Table S4. Genome sequencing and assembly statistics 12

Table S5. Chromosome pseudomolecule statistics 13

Table S6. Sequencing and assembly statistics for RNA-seq libraries 14

Table S7. Summary of Repetitive Elements found in the *C. avellana* genome 18

Table S16. *MLO*-related gene model predictions and results of re-annotation 26

List of Figures

Figure S1. Comparison of SSR marker order on chromosomes with genetic maps. 15-16

Figure S2. Relative abundance of interspersed repeat families in the *C. avellana* genome. 19

Figure S4. Genome statistics of non-coding RNA elements in *C. avellana.* 21

Figure S5. Dot plot representing whole genome alignment of the *C. avellana* and *B. pendula* chromosome assemblies. 24

Figure S6. Multiple sequence alignment (1-540 aa) of MLO protein sequences 27-28

Figure S7. Phylogenetic inference (ML) of selected tree *MLO*-like genes. 30

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Supplementary Methods

1.1 DNA and RNA purification

Genomic DNA was isolated from hazelnut tissues following the method of Wang et al. (2013)[1] with the following modifications: the incubation time with 2x CTAB buffer was shortened to 1 hr at 65^{o}C, and RNAse A was applied in this step rather than as a separate incubation. After resuspension in high-salt TE buffer (1 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA) DNA was bound to a silica membrane column (NucleoSpin Plant II Kit, Macherey-Nagel, Düren, Germany) and washed with 70% ethanol following the kit manufacturer's instructions, to remove low molecular weight DNA fragments, before being eluted in 60 µl of TE buffer.

The same protocol was modified for RNA isolation. All buffers were newly prepared using nuclease-free water. Fresh buds were picked, flash frozen in liquid nitrogen, and ground using 3 mm steel beads in a Tissue Lyser II (Qiagen, Santa Clara, CA). During the 2 x CTAB incubation, RNAse A was not applied; nucleic acids were subsequently recovered by extraction with 24:1 chloroform:isoamyl alcohol. The RNA-containing aqueous phase from this extraction was precipitated using isopropanol. The pellet was cleaned by resuspension first in high salt TE buffer, then in TE without added salt, with ethanol precipitation after each step. The final total RNA was resuspended in 60 µl TE and residual DNA eliminated using RNAse-free DNAse I. In order to confirm that the RNA was not degraded and of sufficient quality for RNA sequencing (RNA-Seq) analysis, 2 aliquots of each sample were also analyzed by agarose gel electrophoresis.

1.2 Sequence data processing and genome assembly

The quality of the raw reads from Illumina and Nanopore sequencing was assessed using FastQC[2].

Illumina only assembly

For Illumina-only assembly, raw sequence reads were processed with Trimmomatic[3] to remove TruSeq adapters, trim bases with quality score <5 from both ends of the reads, and use a sliding window of 4 bases to cut the reads when average sequence quality across the window dropped below 20. The trimmed sequences were then assembled using ABySS 1.9[4] with a range of k-mer size values; `k=80` was found to empirically to give the most contiguous assembly, which was improved further by enabling scaffolding across large bubbles in the k-mer graph (`POPBUBBLES_OPTIONS=--scaffold b=5000`).

Hybrid assemblies

MinION sequencing adapters were trimmed from the first and last 50 nt of NanoPore reads using bbduk from the BBtools suite[5] with the options k=19, editdistance=3. Hybrid genome assembly of the Illumina and trimmed NanoPore reads was carried out using MaSuRCA 3.2[6], with the average Illumina insert size specified as 790±80 nt. This assembly took approximately 9 days running on 36 CPUs in parallel.

Proximity ligation data was integrated with both the Illumina-only and the Hybrid genome assembly using Dovetail Genomics' HiRise assembly pipeline.

Transcriptome mapping

Raw RNA-seq reads were processed with Trimmomatic as described above. Trimmed reads were aligned to the genome using HISAT2 2.1.0[7], using recommended alignment parameters for distinguishing between genotypic variants. StringTie 1.3.4[8] was employed to assemble the transcripts from RNA-seq reads aligned to the genome, and compute their abundance.

1.3 Chromosome structure verification & map comparison

The SSR-based genetic map for OSU 252.146 x OSU 414.062 [9] was compared with the chromosomes as follows: firstly, all SSRs within the chromosome sequences were identified, and each locus extracted along with 500 nt of upstream and downstream sequence using SciRoKo 3.2[10]. Next, PCR primer sequences for 290 previously mapped SSR markers were

collated[9,11,12]. The SSR flanks were searched for these primer sequences using blastn (-task blastn-short) and requiring >95% sequence identity. When both primers matched the flanks of the same SSR locus, this was taken as a positive match, and the chromosome position of the locus compared with the genetic map position of the same SSR (Supplementary Figure 1). Genetic map/chromosome comparisons were drawn using MapChart 3.2[13].

1.4 Telomeric and Centromeric repeats

Chromosome pseudomolecules were searched for classical plant telomere repeats using BLASTN with $(GGGATTT)_5$ as the query seqeunce. In order to identify putative centromeric regions, the chromosomes were searched using tblastn for integrase domain protein sequences specific to centromere-associated CRM class LTRs from *Zea mays* and *Hordeum cereba* (Genbank Accession Nos: AY129008 & AY040832). Matches with < 60% sequence identity or < 80% query coverage were eliminated.

1.5 Discovery and masking of repetitive elements

For identification of candidate novel retrotransposons, the following repeat-type specific prediction tools were used: LTRharvest/LTRdigest for LTRs and TRIMs[14,15], SINE-finder for SINEs[16], TRF for arrays of short tandem repeats[17]. In order to identify novel LTR elements, Hidden Markov Models (HMMs) of known LTR proteins were downloaded from Pfam (Table S1) and supplied as a search library for LTRdigest. Novel LTR elements were clustered using Vmatch[18] following the 80-80-80 principle (80% sequence identity with over 80% coverage of the shorter element, and > 80 bp). After selecting one representative of each cluster, degenerate copies were identified using blastn searches separately for the 5' terminal repeat, 3' terminal repeat, and the internal region of each element. Those with at least 5 LTR copies, or at least 3 copies of the internal region, were accepted as novel LTRs.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Table S1. PFAM Protein domains used for LTR identification.

Pfam ID	Pfam name	Description
PF07253	Gypsy	Gypsy protein
PF00077	RVP	Retroviral aspartyl protease
PF08284	RVP_2	Retroviral aspartyl protease
PF00078	RVT_1	Reverse transcriptase
PF07727	RVT_2	Reverse transcriptase
PF06817	RVT_thumb	Reverse transcriptase thumb domain
PF06815	RVT_connect	Reverse transcriptase connection
PF00075	Rnase H	Ribonuclease H domain
PF02022	Integrase_Zn	Integrase Zinc binding domain
PF00665	rve	Integrase core domain
PF00098	zf-CCHC	Zinc knuckle domain
PF00385	Chromo	'Chromo' (CHRomatin Organisation MOdifier) domain
PF01393	Chromo_shadow	Chromo shadow domain
PF00692	dUTPase	dUTPase domain
PF03078	ATHILA	ATHILA ORF-1 family
PF04195	Transposase_28	Putative Gypsy type transposon
PF05380	Peptidase_A17	Pao retrotransposon peptidase
PF01577	Peptidase_S30	Potyvirus petidase
PF13975	gag-asp_protease	Retroviral aspartyl protease
PF13976	gag_pre-integrs	Retroviral Gag
PF03732	Retrotrans_gag	LTR gag (Gypsy type)
PF14223	Retroran_gag_2	LTR gag (Copia type)
PF18103	SH3_11	Retroviral Int
PF00552	IN_DBD_C	Retroviral Int

These newly identified repeats were combined with known eudicot repetitive elements from RepBase Update 22.08 (5,913 elements) and mipsREdat 9.3 (26,123 elements) to create a custom repeat library. This was then used to search the genome by RepeatMasker 4.0.7 on 'normal' sensitivity with default scoring matrices. Detected repeats were masked with runs of 'N'. For subsequent identification of protein-coding genes and other functional elements, the `-nolow` option was used to leave simple repeats and low-complexity regions unmasked.

1.6 Prediction of hazelnut tRNA, rRNA and miRNAs

Genes encoding tRNAs were identified using tRNAscan-SE 2.0.0[19].

To find rRNA genes, the masked chromosome sequences were searched using BLASTN (e-value cut-off 1e-30) with the following coding sequences: previously published 5S rRNA of *C. avellana* (Genbank HF542974.1[20]); and complete tree 45S rRNA sequences retrieved from Genbank. The pseudochromosomes were searched for sequences with similarity to known ribosomal proteins from other plants recorded in NCBI (listed in Table S9) using BLASTN. Genome regions passing a match threshold (>80 % identity, >80 % query coverage and >200 bp in length) were selected, and putative protein sequences searched using BLASTX (e-value cutoff 1e-6) against the NCBI non-redundant plant protein database. Sequences with a closest match to non-ribosomal proteins were removed.

Conserved miRNA genes were identified using all reported plant miRNAs from miRBase v22 [21] with SUmirFind and SUmirFold[22], using a mismatch cutoff ≤ 3 for initial miRNA homolog detection, followed by predicted pre-miRNA secondary structure prediction and selection of strong miRNA candidates based on established structural criteria[22]. For inter-species comparison, the same methods were used to identify miRNA genes in other plant genomes. Expression of putative miRNAs was confirmed by using all non-redundant pre-miRNA sequences to search assembled Tombul transcripts using BLASTN, retaining hits with >78% identity, >80% coverage as expressed pre-miRNAs. Experimentally validated targets of specific miRNA families found in *C. avellana* were retrieved from miRBase. Predicted targets of putative miRNAs were found in the transcripts using psRNATarget[23], and annotated with Gene Ontology terms using Blast2GO software[24].

1.7 Construction and annotation of gene models

Augustus software was used to generate *ab initio* gene predictions from the masked Tombul genome based on Hidden Markov Models optimised for *Arabidopsis thaliana* [25].

Evidence for gene expression was produced by aligning the gene models to the Tombul transcriptome using BLASTN, retaining only the best hit for each gene model. Conserved orthologs of known plant genes were identified using the 'Assign proteins to groups' pipeline of OrthoMCL-DB (https://orthomcl.org/orthomcl/)[26]. Multiple BLAST hits between different segments of the same pair of sequences were merged using a custom Perl script to calculate the 'Cumulative Identity Percentage' (CIP) and 'Cumulative Alignment Length Percentage' (CALP)[27]. Low-quality hits were eliminated based on the following criteria: CIP < 90%, CALP <60% for transcript matches, and CIP < 70%, CALP < 60% for ortholog matches. Gene models were selected as 'high-confidence' if they were supported by a good quality match to a transcript sequence and/or to an orholog from other plant species. The OrthoMCL analysis also identified 12,516 genes that formed clusters containing only gene models from *C. avellana,* with no orthologs from the other plant species in OrthoMCL-DB. These putative hazelnut-specific genes were added to the high-confidence list if: i) they had a transcript match at CIP >90% and CALP >90%, showing that they are expressed; or ii) they matched an expressed high-confidence gene at CIP > 70% and CALP > 60%, suggesting that they could be paralogs resulting from gene duplication. This gave a working set of 27403 gene models. Automated functional annotation of the gene models was carried out using Mercator4, and TRAPID2 with default parameters[28,29]. Gene models annotated as retroviral/retrotransposon proteins (Reverse Transcriptase, retrotransposon Gag, chromodomain) or assigned to OGs corresponding to these proteins were investigated (Table S1), and gene models that lacked transcript support were eliminated as probable degenerate repeats[30]. The final edited gene set contained 27270 gene models.

Table S2. Ortholog Groups containing transposable element domains in the *C. avellana* gene set.

Group ID from OrthoMCL-DB[31]	Domain description	No. of orthologs in *C. avellana* gene set	No. of orthologs with transcript evidence
OG5_126614	Reverse transcriptase	3	0
OG5_131059	MULE transposase	8	5
OG5_131652	Reverse transcriptase	1	0
OG5_138969	Retrotransposon gag protein	3	0
OG5_126614	Reverse transcriptase	6	0
OG5_153395:	Plant Mobile Domain	2	2
OG5_156509:	Plant transposase (Ptta/EnSpm)	0	0
OG5_159949:	Reverse transcriptase	2	0
OG5_184563:	Plant transposase (Ptta/EnSpm)	0	0
OG5_212464:	Plant transposase (Ptta/EnSpm)	0	0

1.8 Data for comparative genomics and phylogenetic tree construction

The final set of high-confidence gene models from *C. avellana* var. 'Tombul' was compared with other trees from the order Fagales for which complete genome assemblies and annotated gene sets are available: silver birch (*Betula pendula*)[32], English walnut (*Juglans regia* cv. 'Chandler')[33], and pedunculate oak (*Quercus robur*)[30]. As an outgroup, two well-characterized fruit tree genomes, apple (*Malus domestica* cv. 'Golden Delicious') and peach (*Prunus persica*) were selected from the Rosaceae[34]. Whole genome sequence comparisons were carried out using LAST v869[35].

1.9 Ortholog identification and sequence comparisons

The complete sets of protein sequences for all annotated genes in these six genomes were downloaded and compared with each other using OrthoFinder 2.3.11, using the default pipeline to find orthogroups and infer rooted gene trees for each group[36]. The automatically generated species tree incorrectly assigned *Q. robur* as the outgroup; therefore the tree was modified to place the root on the branch containing the divergence between *Q. robur* (and

the other Fagales) and the Rosales. The final step of the OrthoFinder analysis, reconciling gene trees with the species tree, was then re-run using the correctly rooted tree.

For genome-wide inference of divergence times, the set of 979 single-copy orthologs (SCOs; genes found as a single copy in all 6 species) was selected. Separate fasta files were produced by OrthoFinder containing the 6 predicted protein sequences for each SCO; seqeunces within each file were aligned with each other using ClustalW[37]. Sequence segments that are only present in 1 or 2 orthologs can skew phylogenetic analysis, including timetrees, so alignments were edited using Gblocks 0.91[38], retaining only sequence blocks that were conserved in at least 4/6 species. 113 SCOs for which <30% of the total sequence length was in conserved blocks were considered as suspect/low quality and were removed from the analysis, leaving 866 edited SCOs for the following steps.

1.10 Species and gene set divergence time estimation

Time-scaled phylogenetic trees were constructed using the MCMCTree program from the PAML 4.9j package (Phylogenetic Analysis by Maximum Likelihood)[39]. MCMCTree carries out Bayesian inference of divergence times on a sequence alignment using a variety of relaxed molecular clock models, and a guide phylogenetic tree including soft fossil calibrations[40]. The edited gene sequence alignments created above were concatenated and converted into sequential Phylip format as input for MCMCTree using catfasta2phyml.pl[41]. Calibration points for the guide tree were selected from previous detailed surveys of fossils from the Fagales[42] and all angiosperms[43] and are summarized in Table S3. For each divergence of the tree, the most recent possible age of the oldest extant fossil(s) within daughter lineages were used to set a hard minimum age for the crown group. The extinct species *Cranea wyomingiensis* was originally described as belonging to the Coryloideae[44]; however it also has morphological characters of other Betulaceae genera, so here was left in the Betulaceae crown group.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Table S3. Fossil calibrations used for calculation of timetrees. The 4th and 5th columns indicate the branch points to which calibrations were added, and the figures used.

Fossil species	Age estimate (mya)	Taxon	Divergence point on timetree	Minimum age (mya)
Betula leopoldae *Alnus* sp.	37.2 - 48.6 48.6 - 55.8	Betuloideae	*Betula* \ *Corylus*	49
Carpinus perryae *Corylus johnsonii*	49 - 50 49 - 50	Coryloideae		
Cranea wyomingiensis	55.8 - 65.5	Betulaceae	*Betula & Corylus* \ *Juglans*	55.8
Cyclocarya brownii *Polyptera manningii*	55.8 - 65.5 56.8 - 66.3	Juglandaceae		
Prunus wutuensis *Oemleria janhartfordae*	48.6 - 55.8 49.4 - 54.2	Rosaceae	*Malus* \ *Prunus*	49
Archaefagaceae futabensis	87.5 - 96.3	Fagales	Rosales \ Fagales	87.5

The MCMCTree analysis was realized using a relaxed clock model with independent rates on each branch of the tree following a log-normal distribution, and the root age constrained to <120 million years ago (mya). The JC69 substitution matrix was used and the fossil ages indicated in Table S3 were set as soft minimum bounds for each node using the default Cauchy probability distribution (see PAML documentation, except that the scale factor was reduced from 1 to 0.5. The Markov chain was allowed to burn-in for 4000 iterations, and samples were then collected every 20 iterations until a total of 20,000 had been collected. The MCMCTree process was run twice and the dating estimates from both runs compared; stable time and rate estimates during the run, and highly consistent values between runs (R^2 > 0.995) were taken to indicate that the Markov chains had converged.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Supplementary Results and Discussion

2.1 Genome sequencing and hybrid assembly approach

As discussed in the main text, 3 different whole-genome shotgun sequencing technologies were used to sequence the *C. avellana* var. 'Tombul' genome; short paired-end reads (Illumina PE), long reads (NanoPore) and proximity ligation sequencing (Dovetail Chicago-HiC). Genome assembly was carried out for Illumina only and for each combination of 2 or 3 of the technologies (Table S4). The 'Illumina only' assembly was highly fragmented and ~25% larger than the expected genome size (378 Mb), probably due to heterozygous regions being assembled twice into separate contigs. A hybrid assembly of the Illumina and NanoPore data (Illumina + NP) gave 12,557 scaffolds that were ~10-fold larger than 'Illumina only' across the size distribution and eliminated the large majority of duplicated sequences and gaps in the assembled scaffolds. However, this assembly was still too fragmented (N90 = 5857) to allow large-scale structural comparisons. In both genome assemblies, the observed GC content was 36%.

Table S4. Genome sequencing and assembly statistics

Sequencing technology	**No. of Reads**	**Library design**	**Total read length (Gb)**	**Sequence coverage[1*]**
Illumina PE	2 x 136M	Paired end, 700-800 bp insert	41.1	108 x
NanoPore	1,351,274	Single molecule, 1-10 kb reads	3.53	9.3 x
			Insert size range	Physical coverage[2*]
Dovetail Chicago	2 x 221M	Proximity ligated	1-100 kb	225 x
Dovetail HiC	2 x 242M	Proximity ligated	10 kb – 10 Mb	3,447 x

Assembly Statistics				
	Illumina only	Illumina + Dovetail	Illumina + NP	Illumina + NP + Dovetail
No. of scaffolds (>1kb)	89,427	32,741	12,557	2,206
Scaffold N50 / L50	20927 / 7.32 kb	12 / 13.33 Mb	1299 / 78.8 kb	5 / 36.65 Mb
Scaffold N90 / L90	149112 / 204 bp	89585 / 3.4 kb	5857 / 13.7 kb	10 / 22.72 Mb
Total scaffold size (Mb)	513.8	520.1	383.1	384.2
% gap bases (N)	6.01%	6.45%	0.18%	0.47%
Largest scaffold	0.152 Mb	45.9 Mb	0.732 Mb	50.95 Mb

[1]Average no. of times a genome nucleotide is included in a sequence read

[2]Average no. of times a genome position is included in the region between 2 linked reads

* Calculated for an estimated genome size of 378 Mbp.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

In order to improve the contiguity of the assembly, we carried out proximity ligation sequencing using Dovetail Genomics' proprietary methods. These generate pairs of linked reads that originate from within the same large DNA fragment (Chicago library) or from physically adjacent nucleosomes in native chromatin (HiC library). The sequences of the linked reads are not assembled directly, but mapped to the scaffolds from the pre-existing genome assembly. The 'HiRise' bioinformatic pipeline then uses these links to determine the order and orientation of scaffolds along each pseudomolecule. Adjacent, non-overlapping scaffolds are joined with an arbitrary gap sequence of $(N)_{100}$. As shown in Table 1, incorporation of Dovetail data enabled >50% of the genome to be assembled into chromosome-sized pseudomolecules both from Illumina-only and Illumina + NP assemblies. However, a very large number of small scaffolds remained unassembled in the 'Illumina + Dovetail' assembly, reflected by the very high N90 (Table S4), and the duplicate scaffolds were not resolved. In contrast, the Illumina + NP + Dovetail assembly consisted of 11 chromosome-sized pseudomolecules ranging from 22.42 – 50.95 Mb in length (Table S5), in total accounting for 97.8% of the predicted genome size. The next largest scaffold after these contained the chloroplast and mitochondrial DNA, after which there remained 2194 unplaced small scaffolds (1 kb - 400 kb) with a cumulative length of 13.6 Mb.

Table S5. Chromosome pseudomolecule statistics

Pseudo-molecule ID	Size (Mbp)	Size group	No. of gaps ('N' percentage)	OSU linkage group (Size in cM)	Matched SSRs (no. / no. colinear)
pchr01	50.95	L	1337 (0.48%)	1 (193)	16 / 16
pchr02	50.86	L	1565 (0.54%)	2 (185)	10 / 10
pchr03	39.77	M	1010 (0.43%)	4 (153)	14 / 12
pchr04	36.85	M	914 (0.47%)	9 (184)	14 / 13
pchr05	36.65	M	1031 (0.46%)	5 (123)	11 / 11
pchr06	30.27	M	816 (0.47%)	7 (123)	9 / 8
pchr07	30.24	M	609 (0.49%)	11 (118)	6 / 6
pchr08	25.77	S	665 (0.46%)	3 (105)	10 / 10
pchr09	23.27	S	465 (0.38%)	10 (101)	8 / 8
pchr10	22.72	S	557 (0.55%)	6 (88)	2 / 2
pchr11	22.42	S	630 (0.49%)	8 (103)	13 / 12
Total	369.78		9599 (0.48%)		113 / 108

2.2 Transcriptome Assembly

Two RNA-seq libraries were generated from total RNA isolated from leaf buds, and were sequenced by following Illumina protocols with results as indicated in Table S5. After quality control and trimming, the reads were aligned to the chromosome pseudomolecules using Hisat2. Aligned reads were assembled using StringTie individually for each sample, and then redundant transcript contigs from both samples were combined, using StringTie's 'merge' function; at this step, putative transcript isoforms with very low expression levels (<1% of all expression for that transcript) The final assembly included 57315 distinct transcripts arising from 47250 transcribed 'genes' (Table S6). This is expected to be greater than the number of protein-coding genes in the genome, because it includes non-coding RNAs; also, heterozygous alleles may be assembled twice by this method, into different 'genes'.

Table S6. Sequencing and assembly statistics for RNA-seq libraries (Paired end, 2 x 100 nt)

Sample ID	**Total read bases (bp)**	**Total reads**	**GC(%)**	**AT(%)**	**Q20(%)**	**Q30(%)**
Library_T1	6,677,685,902	66,115,702	46.310	53.69	98.763	95.400
Library_T2	7,017,024,288	69,475,488	46.084	53.92	98.163	95.286
Mapping & assembly	Filtered reads	Reads mapped & paired	% Properly paired	Assembled transcript contigs	Merged isoforms	Final 'genes'
Library_T1	65,937,538	61,974,596	94.0	76540	57315	47250
Library_T2	69,238,960	62,411,216	90.1	77602		

2.3 Chromosome structure verification & map comparison

The chromosome arm structure of the Tombul pseudomolecules was verified by comparison with existing genetic maps (Supplementary Figure 1) and identification of telomeric and centromeric repeats. Telomeres are difficult to assemble due to their highly repetitive nature; indeed classical plant telomere repeats $(GGGATTT)_n$ were only observed at the ends of 4 of the chromosome pseudomolecules, while a possible non-canonical telomere sequence (GGGGTTTCATTT) was found at the ends of 2 further pseudomolecules.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

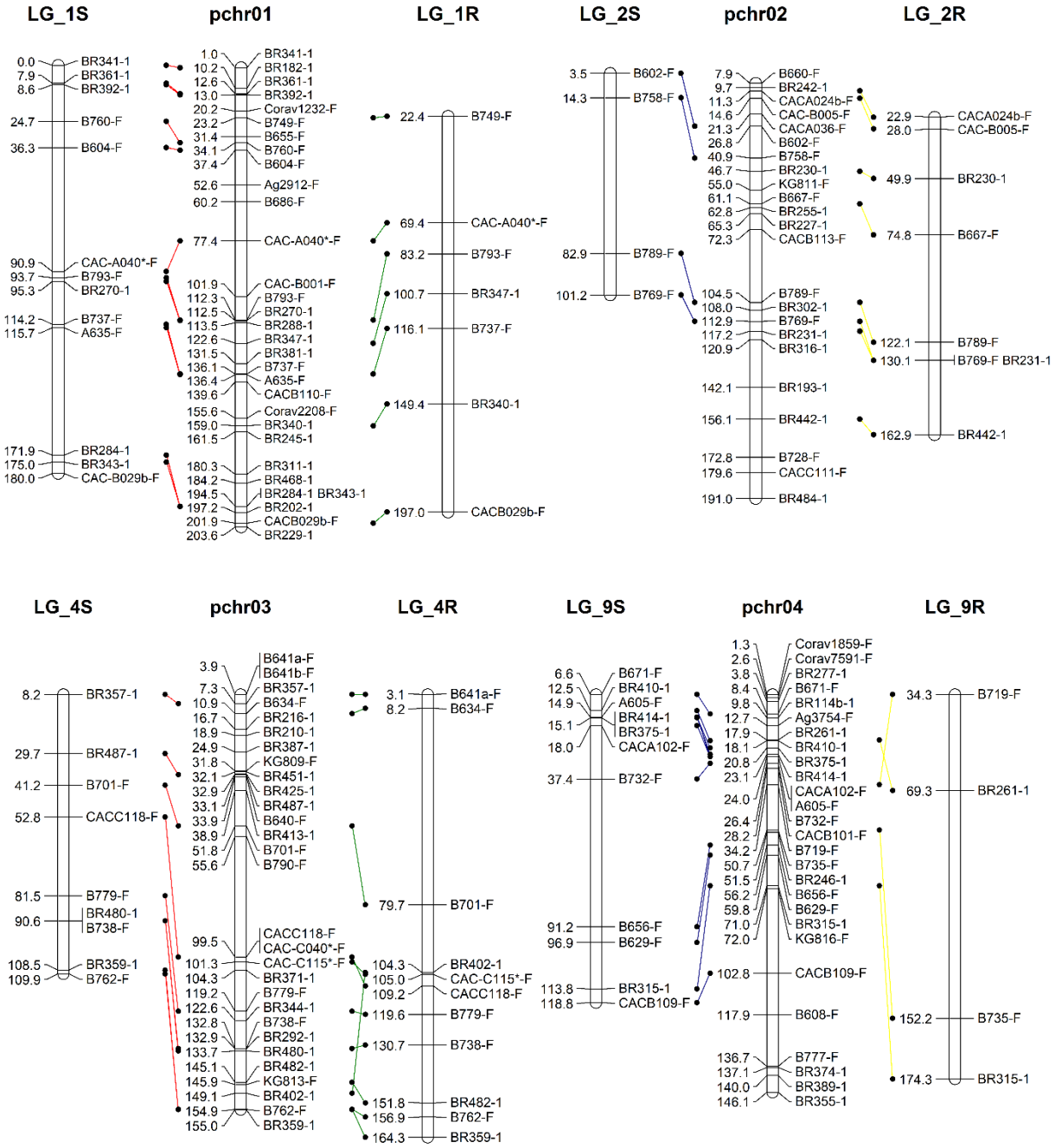
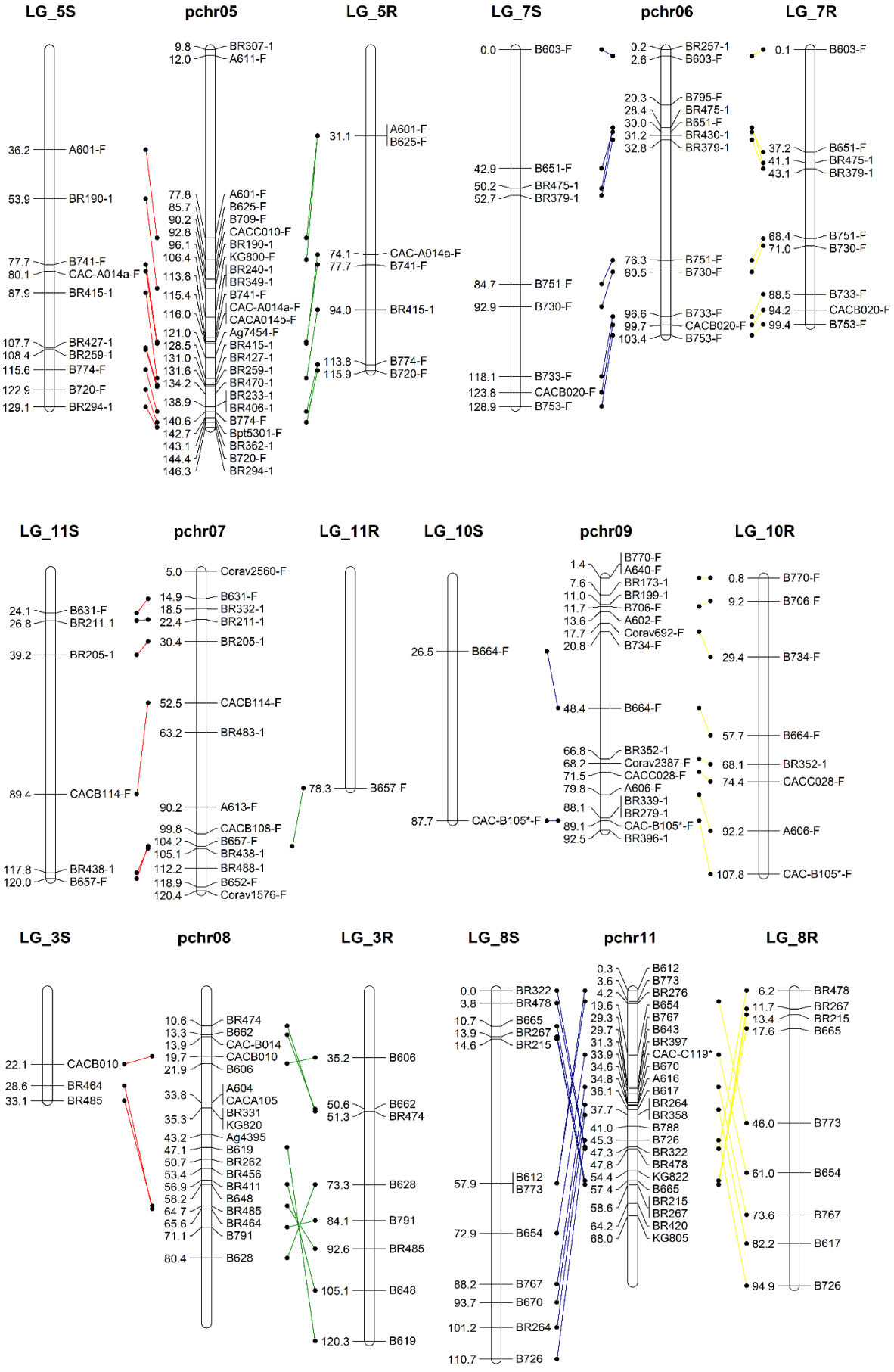


Figure S1. Comparison of SSR marker order on chromosomes with genetic maps. LG_1S/1R indicate the linkage groups from the OSU 252.146 x OSU 414.062 cross, with marker positions marked in cM, and only showing markers also found on the chromosomes. To give a similar scale, the markers on the chromosome bars are labelled with their bp position/250,000. The orientation of markers on mis-assembled chromosomes (next page) showed the necessary correction; in pchr08, the chromosome arms are inverted; in pchr11, they are correctly orientated but the long arm is stacked above the short arm. The map for LG_6S/6R & pchr10 is not shown owing to the small number (2) of shared markers.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)



The CRM sub-family of Gypsy LTR retrotransposons has a uniquely conserved integrase domain (the 'chromodomain') and is associated with peri-centromeric regions in plants[45]. Searching for this domain found 2-3 copies near the centre of 8 of the chromosomes, while no high-confidence match was found on pchr02. Notably, the 2 remaining pseudomolecules (pchr08 & pchr11) appeared to have CRM family repeats at both ends and telomeric repeats near the centre, suggesting that the chromosome arms were erroneously joined at the telomeres rather than the centromere.

Comparison with the published genetic map (Supplementary Figure 1, bottom) confirmed this; in these two pseudomolecules, the SSRs mapped to the halves of each chromosome were co-linear with those in the corresponding linkage group, but the order of the 2 arms was reversed. In both cases the false join was determined to have been made at the final assembly step (Dovetail HiC) where two scaffolds that both had classical telomere repeats at one end were joined together. Accordingly, these two chromosome sequences were manually split at the telomeric repeats, the halves inverted and re-joined according to the order of SSR markers shown by the genetic map.

This assembly error can be explained as an artifact of the HiC ligation method; if the two ends of the chromosome are located in close physical proximity in native chromatin, links could be created between the two sub-telomeric regions. Similarly, in native chromatin few links may be produced between the highly condensed peri-centromeric regions, even though they are adjacent. In these circumstances, the HiRise pipeline could mistakenly join the chromosome arms at the telomeres rather than the centromere.

After rearrangement, 108 of 113 SSRs shared between the chromosomes and the OSU genetic linkage map were co-linear.

2.4 Detection and Classification of Repetitive elements in the *C. avellana* genome

About a third of the Tombul genome was found to consist of LTR retrotransposons (summarized in Table S7), the majority of which were copies of 598 novel LTR elements identified in this study (Supplementary Methods). New Gypsy & Copia superfamily elements were distinguished by the order of protein-coding domains in their internal regions[46]. When insufficient protein-coding domains were present positively to identify the family, the elements were classified simply as 'LTR'.

Table S7. Summary of Repetitive Elements found in the *C. avellana* genome

	Number of elements	**Length occupied (bp)**	**Percentage of sequence (%)**
Retroelements	350,766	131,114,230	35.38
SINEs:	239	14,731	0
LINEs:	620	103,692	0.03
- RTE/Bov-B	96	19,947	0.01
- L1/CIN4	551	99,487	0.03
LTR elements:	349,907	130,995,807	35.35
- Ty1/Copia	101,052	32,516,429	8.78
- Ty3/Gypsy/DIRS1	50,068	25,164,581	6.79
Retroviral	0	0	0
DNA transposons	2,397	1,123,336	0.3
hobo-Activator	787	354,147	0.1
En-Spm/CACTA	385	164,418	0.04
MuDR-IS905	438	182,472	0.05
Tourist/Harbinger	234	98,731	0.03
Helitron	488	269,216	0.07
Unclassified:	589	106,975	0.03
Total interspersed repeats:		**132,344,541**	**35.72**
Small RNA:	399	41,681	0.01
Satellites:	7	579	0
Other simple repeats:	208820	7,134,639	1.93
Low complexity:	36677	1,756,537	0.47
Total simple repeats:		**8,933,436**	**2.41**
Total repetitive content:		**141,262,718**	**38.13**

TRIMs were identified using the following criteria[47]: i) having terminal repeats and 4-6 bp target site duplications, simlar to LTRs, ii) maximum length 1500 bp, iii) containing no retrotransposon protein coding genes, and iv) having at least 10 copies in the genome. Putative novel SINEs were required to have >60% sequence identity over at least 60 bp to a known tRNA gene, and multiple full-length copies in the genome[48].

In total the chromosome pseudomolecules were found to consist of 35.72% interspersed repeats, and 2.41% simple repeats (Table S6, Figure S2).

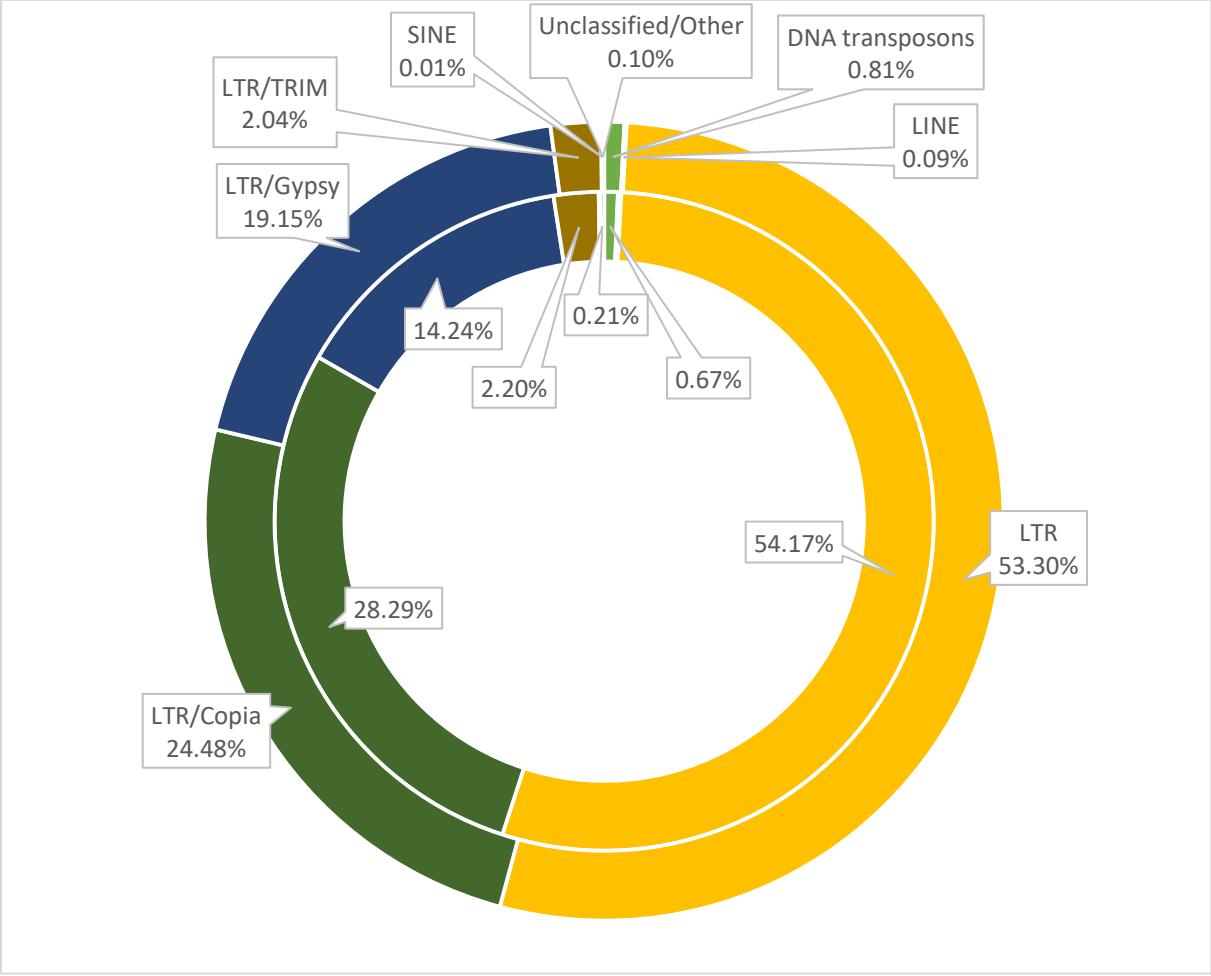


Figure S2. Relative abundance of interspersed repeat families in the *C. avellana* genome. Outer ring shows percentage of total sequence coverage by interspersed repeats (132.3 Mb from the whole genome); inner ring shows percentage of total number of interspersed repeat loci. LINEs and SINEs are not indicated on the inner ring due to their low abundance.

The overall abundance of repetitive elements is similar to the rice genome (35-40% depending on genome assembly[49]) but lower than that of the slightly larger silver birch genome (*Betula pendula* L., 51.35%[32]). Over half of the LTR elements were too diversified to be positively assigned to a sub-family; of the rest, Copia elements were almost twice as abundant as Gypsy elements, but owing to their shorter average length, the total number of bases covered by both families was comparable. TRIMs (Terminal Repeats In Miniature) are a relatively recently described class of short, non-autonomous LTR retroelements, that are thought to replicate using the enzymes encoded by related Copia or Gypsy LTRs[47]. TRIMs were found to be highly abundant in the *B. pendula* genome with over 82,000 copies, but their copy number in *C. avellana* was only 8,712, which is within the typical range observed for other plant genomes. Terminal inverted repeat (TIR) DNA transposons, which were again abundant in *B. pendula,* were notably scarce in *C. avellana*; consistent with this, there were no intact copies of the EnSpm/Ptta transposase, responsible for replicating the TIR/CACTA transposon family, detected in the genome (Table S2). These observations suggest that the expansion of these repeat families in *Betula* took place after the ancestors of the *Betula* and *Corylus* lineages diverged.

2.5 Annotation and inter-species comparison of tRNA genes

A total of 477 predicted tRNA genes and 40 tRNA pseudogenes, with an average length of 75 nt, are listed in Table S7. Of these 27 tRNA genes were predicted to contain introns, chiefly some of those encoding Gly, Met & Tyr. Codon frequencies for tRNA genes in *C.avellana* were calculated and compared with other tree species and Arabidopsis (Table S8). tRNA predictions for *A. thaliana* were taken from GtRNAdb[50].

Various differences in abundance between tRNA genes for particular amino acids, and codons for the same amino acid, were observed (Supplementary Figure 3); for example, tRNAs for codons ending "–AT" were non-existent or rare in all the plant genomes examined. Some species-specific differences were also noted; *C.avellana* has a higher abundance of tRNAs for

all Glycine codons, and especially GGC, than the other plant genomes, whereas *P. trichocarpa* has an increased number of all Alanine & Leucine tRNAs. Among the Serine codons, *C. avellana* has a greater abundance of tRNAs for AGC than the other tree species, whereas Arabidopsis favours TCT. Further study would be useful to determine whether these differences are correlated to codon frequencies in protein-coding genes, or influence translation efficiency in different plant species; however, overall there were few species- or lineage-specific variations in tRNA abundance.

2.6 Ribosomal RNA and protein distribution

The only full-length 45S rDNA was identified on pchr08, however, there was also a partial rDNA locus on pchr02. 21 copies of the 5S rDNA were identified on 7 different chromosomes, with the largest cluster of 8 tandemly repeated copies on the short arm of pchr05, followed by a cluster of 3 copies on pchr11 (Table S9). Homologs of ribosomal proteins from other plant species were also found distributed across most of the chromosomes, but absent from pchr11 (Figure S4A & Table S9). In contrast, tRNA genes were relatively evenly spread among the chromosomes, while miRNA families were largely absent on pchr09, suggesting that these different classes of non-coding RNA elements have undergone independent assortment and expansion during the evolution of the *C. avellana* genome.

2.7 MicroRNA annotation and target identification

153 candidate conserved miRNA genes were predicted in the *C. avellana* genome based on previously established annotation criteria (Table S7). The copy number of conserved miRNA families was determined by removing redundant matches from multiple query miRNAs in the output of SUmirFold, to identifying unique miRNA-coding sequences. In some cases, several mature miRNAs from other plants matched the same pre-miRNA candidate; for example, a miR171 family gene located from 944161-944280 on pchr01 had 3 possible mature miRNA predictions from different parts of the pre-miRNA. In these cases all candidates were retained

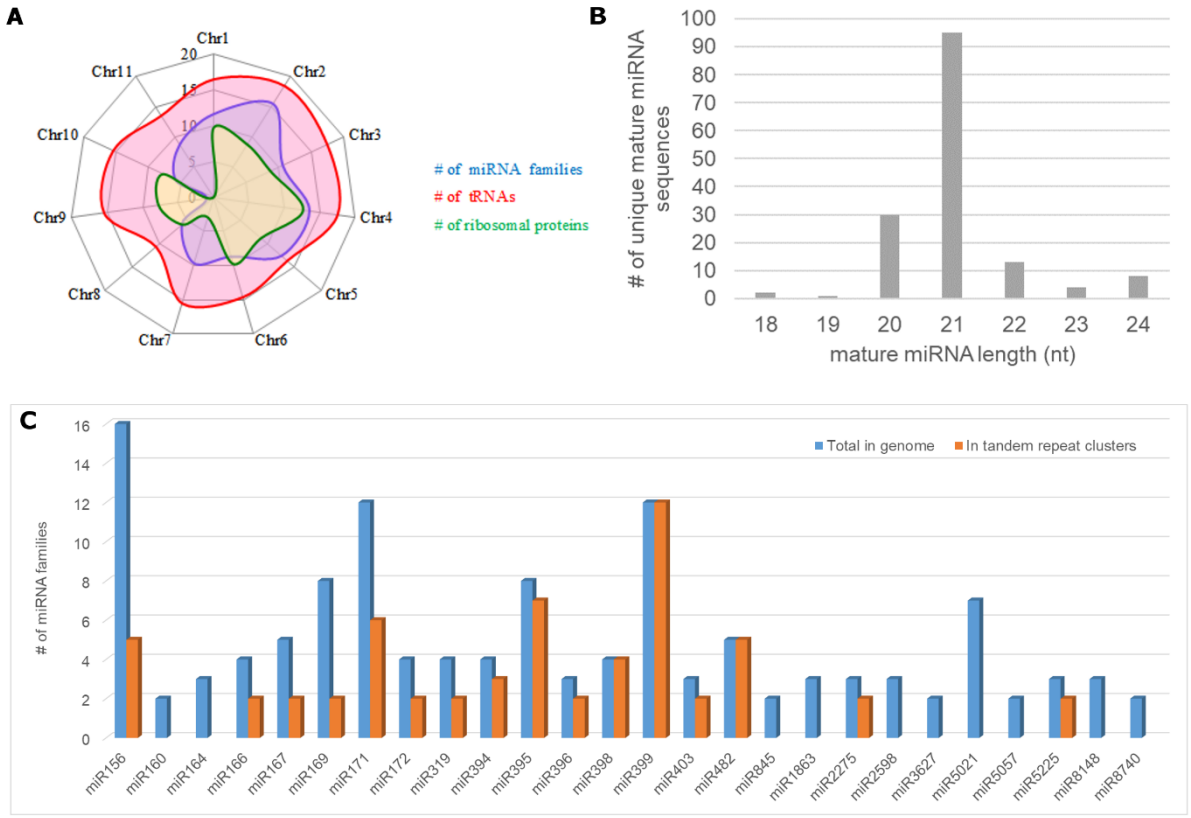


Figure S4. Genome statistics of non-coding RNA elements in *C. avellana.* **A.** Radar plot of the distribution of ribosomal proteins, tRNA genes and miRNA families on *C. avellana* cv Tombul chromosomes. **B.** Distribution of predicted mature miRNA lengths. **C.** Copy numbers of miRNA families, in the genome and in tandem repeat clusters.

in the table, but experimental confirmation, e.g. from small RNA sequencing, should be carried out to confirm which are produced *in vivo*. Predicted mature miRNAs ranged from 18-24 nt in length with the majority being 21 nt (Fig. S4B). The 153 candidate miRNA genes included members of 52 miRNA families, of which 26 had multiple copies in the genome (Supplementary Fig. 4C). Tandem duplications contributed to the expansion of these miRNA families to a variable degree; for example, most miR171 genes were found as single loci distributed between 5 different chromosomes, whereas all 12 miR399 genes were found in a single cluster on pchr02. These structural variations may provide clues as to the contrasting development

of these miRNA families during the evolution of *Corylus*, and also to their current regulation and functions.

Potential miRNA targets were predicted based on the presence of their complementary target sequences in Tombul transcriptome sequences, which were then annotated with GO terms (Table S11).

2.8 Protein-coding genes and annotation

Prediction of protein-coding genes and annotation with MapMan bins, GO terms and conserved protein domains was carried out as described in the main text, and results are summarized in Supplementary Tables 12+. With respect to the 3 main GO categories, Biological Process GO terms were assigned to 15,425, Cellular Component terms to 15,313, and Molecular Function terms to 16,473 gene models. For each gene model, the most specific (highest level) GO terms assigned for each category are recorded in Supplementary Table 12. The most abundant conserved InterPro protein signatures were derived from a variety of protein kinases; e.g. 1401 gene models were annotated with IPR000719, 'protein kinase catalytic domain', accounting for almost 5% of all high-confidence gene models; this is similar to the abundance observed in other sequenced plant genomes [51] and indicates the importance of signal transduction in controlling many aspects of plant biology. Apart from highly prevalent domain structures that are found in proteins with diverse functions (e.g. NAD(P)-binding, Pentatricopeptide repeat, Zinc finger, WD40, and Armadillo domains) other notable abundant domains included the NB-ARC motif, found in 279 gene models, the large majority of which also contained leucine-rich repeats (LRRs). These genes belong to the NB-LRR family of pattern recognition receptors, which play a critical role in the plant response to pathogens. Homeobox-like and Cytochrome P450 domains were each found in over 300 gene models, making these the most abundant transcription factor and enzyme families respectively in the hazelnut genome.

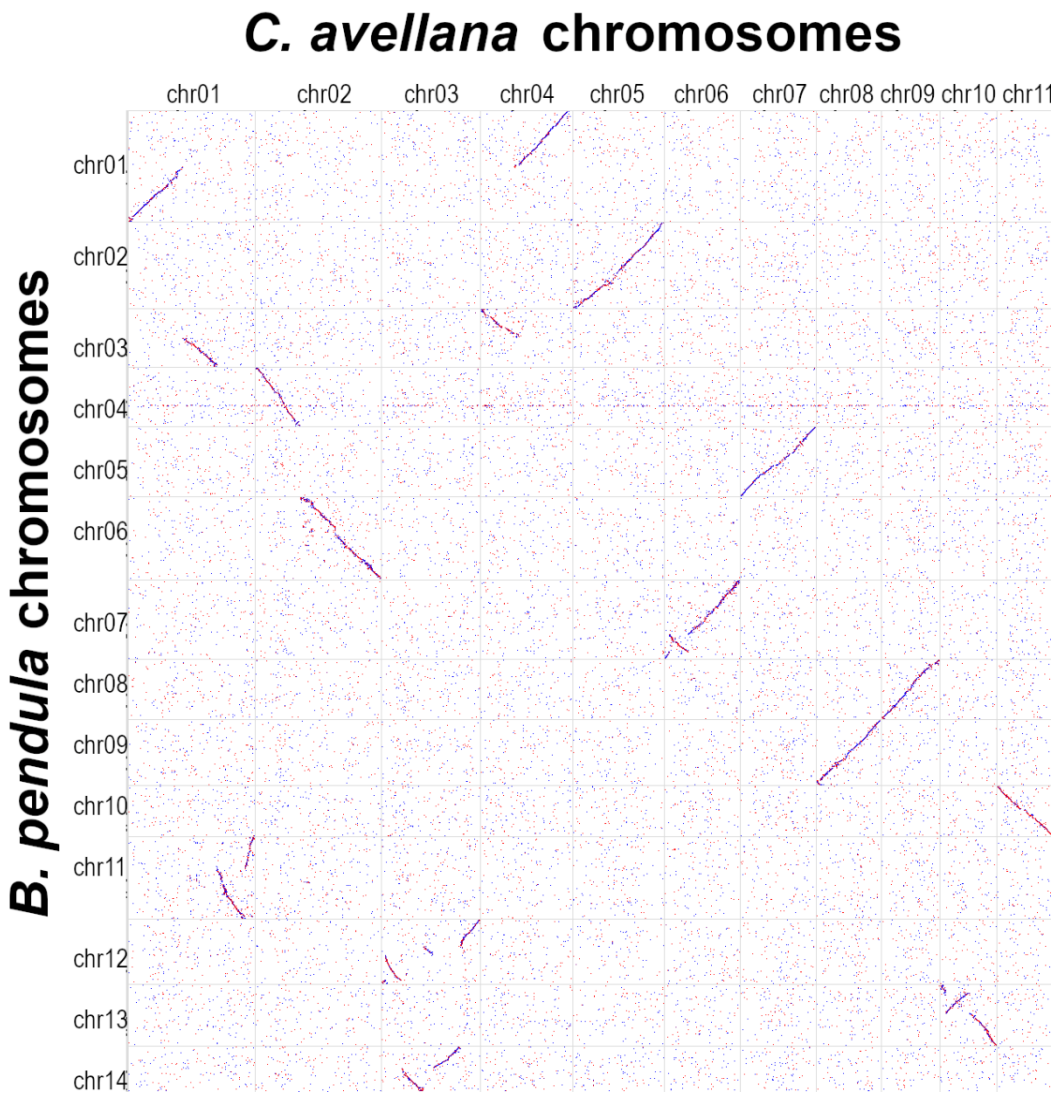


Figure S5. Dot plot representing whole genome alignment of the *C. avellana* and *B. pendula* chromosome assemblies. Blue dots indicate sequence matches in the same orientation, red dots indicate inverted sequence matches.

2.9 Comparative Genome Alignment

The chromosome pseudomolecule sequences from *C. avellana* var. Tombul were aligned with those of *B. pendula*[32] using LAST, with masking of short repeats and other parameters as recommended by the software authors (https://github.com/mcfrith/last-genome-alignments). A dot-plot of the alignment is shown in Figure S5. Strikingly, *C. avellana* pchr05, 07, 08, 09 & 11 are collinear with *B.pendula* chromosomes 2, 5, 9, 8 & 10 respectively along their entire length. pchr06 and pchr10 are largely collinear with *B. pendula* chromosomes 7 & 13 respectively,

with a single large inversion in each. The greatest structural difference between the species is seen in *C.avellana* pchr03, which includes 6 large segments that are rearranged into *B. pendula* chr12 & chr14. pchr02 is split into *B. pendula* chr4 & chr6, while pchr01 & 04 are split into 6 large segments in total, pairs of which are assorted into *B. pendula* chromosomes 1, 3 & 11. The observed structural differences can be explained parsimoniously by assuming an ancestral Betualceae karyotype of 17 chromosomes, with 6 and 3 chromosome fusion events taking place in the *Corylus* and *Betula* lineages respectively, along with 6 intra-chromosomal inversions distributed between the lineages.

2.10 Characterization of the *MLO* gene family

MLO proteins are located at the plasma membrane and have a well-conserved structure consisting of 7 transmembrane (TM) helices, followed by a C-terminal cytoplasmic Calmodulin binding domain, making them relatively easy to identify by sequence homology. Based on the initial automated gene annotation described above, 24 gene models were annotated with the MLO-related protein domain (IPR004326). However, over half of these gene models had a predicted protein product that was significantly shorter or longer than known MLO proteins (<350 or >650 amino acids). Therefore all of the gene models were manually re-annotated, using blastp against non-redundant plant proteins confirm their identity, tblastn to identify regions of MLO homology that had been omitted from gene models, and FGENESH[52] to re-evaluate the gene models predicted by Augustus[25]. Where FGENESH and Augustus gave conflicting predictions, the helical structure of both predicted proteins was assessed using TMHMM[53], and the one most consistent with known MLO protein topology selected. The results of re-annotation are summarized in Table S16. After re-annotation, 11 gene models were retained as encoding full-length MLO proteins; the remaining genes mostly seem to be truncated or disrupted duplicates of the full-length *MLO*s, and so were labelled as '*MLO*-like genes'.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Table S16: *MLO*-related gene model predictions and results of re-annotation.

Gene Model ID	Length (aa)	Best blastp match in *Arabidopsis*	Result of re-annotation
Cav01g18310	140	AtMLO4	Merged with upstream, un-annotated ORF to give full-length *MLO* gene
Cav02g00450	573	AtMLO10	Verified as full-length *MLO* gene
Cav02g01800	306	AtMLO4	Two separate, truncated genes, probably a degraded ancestral *MLO* gene
Cav02g01810	202		
Cav02g28680	484	AtMLO12	Original gene model had only 422 amino acids; replaced with better FGENESH gene model
Cav02g28690	158	AtMLO12	Truncated ortholog of Cav02g28680
Cav02g28700	168	AtMLO6	Truncated ortholog of Cav02g28680
Cav02g28710	180	AtMLO12	Truncated ortholog of Cav02g28680
Cav03g23960	586	AtMLO11	Original gene model had only 537 amino acids; replaced with better FGENESH gene model
Cav04g11090	552	AtMLO1	Verified as full-length *MLO* gene
Cav04g11100	272	AtMLO1	Fragmented ortholog of Cav04g11090, disrupted by an LTR/retroviral insertion
Cav04g11110	247		
Cav07g08600	518	AtMLO1	Verified as full-length *MLO* gene
Cav07g19540	590	AtMLO6	Verified as full-length *MLO* gene
Cav07g19550	348	AtMLO6	Truncated ortholog of Cav07g19540
Cav07g19560	603	AtMLO6	Full-length *MLO,* near ortholog of Cav07g19540
Cav07g19570	106	AtMLO6	Truncated ortholog of Cav07g19540
Cav07g19580	798	AtMLO2	Possibly 2 ancestral *MLO* genes spliced together
Cav07g19590	541	AtMLO12	Original gene model had only 474 amino acids; replaced with better FGENESH gene model
Cav08g17470	403	AtMLO12 (part)	Truncated ortholog, similar to fragment upstream of Cav02g00450
Cav08g17510	520	AtMLO5	7-8 TM helices, closest to Cav02g00450
Cav09g12130	532	AtMLO12	Truncated ortholog (of Cav10g09700?)
Cav09g12140	612	AtMLO3	Original gene model had only 345 amino acids; replaced with better FGENESH gene model
Cav10g09690	427	AtMLO12	*MLO*-like, N-terminal truncated
Cav10g09700	589	AtMLO2	FGENESH gene model, very slightly better than original prediction (587 amino acids)
Cav10g09710	192	AtMLO3	Two separate, truncated genes, probably a degraded ancestral *MLO* gene
Cav10g09720	318		

Multiple sequence alignment of the *C. avellana* MLO genes with those from *A. thaliana* & *Malus domestica* (Figure S6) shows that the TM helix sequences are usually highly conserved, with some insertions in individual genes in the loops between them, and greater diversity at the C-terminus. However, Cav07g08600 seems highly divergent from the other genes in part of the TM region (between 600-750 in Fig. S6); this gene was also only predicted to contain 5 TM

helices instead of 7. Therefore further structural or functional studies are required to confirm whether it encodes an intact MLO protein.

In addition to the alignment of full-length genes mentioned above, full-length *MLO* and *MLO*-like genes from Clades V, VI & VII were aligned with orthologs from the other tree genomes used for evolutionary analysis. Phylogeny within these groups was reconstructed using the Maximum Likelihood method, allowing variable substitution rates between sites (Figure S7). In cases in which an historic *MLO* gene has been degraded into 2 smaller putative genes, the amino acid sequences of the 2 fragments were merged with a 'XXXXX' linker in between, to provide a more complete alignment (e.g. Cav02g28690 & Cav02g28700, Fig. S7A). Based on this reconstruction, while most other *MLO* genes are evenly represented in the tree genomes examined here, both Cav02g28680 (Fig. S7A) and Cav07g19540 (Fig. S7B) appear to have undergone repeated gene duplications that are unique to the *Corylus* lineage.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

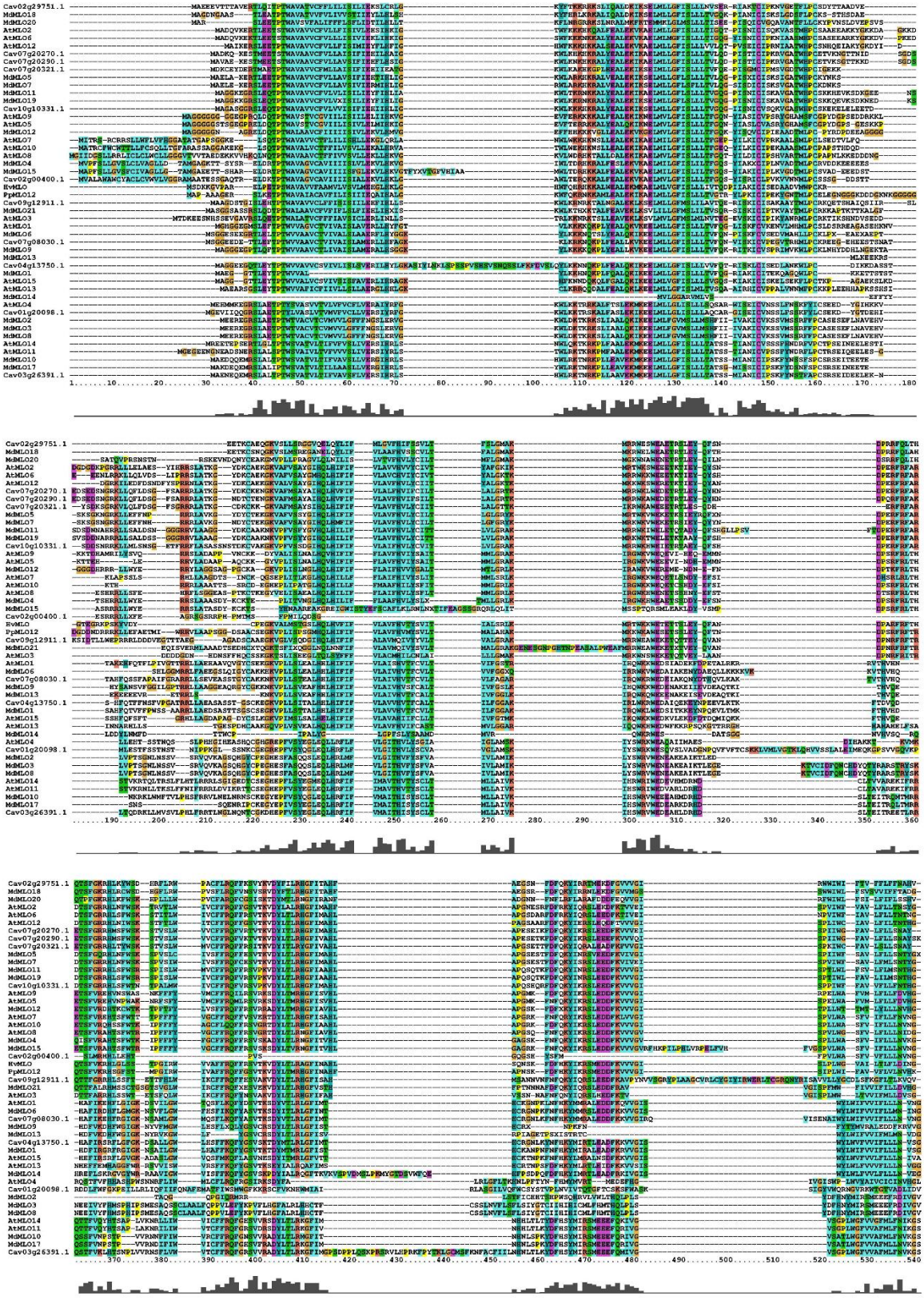


Figure S6. Multiple sequence alignment (1-540 aa) of predicted MLO protein sequences from *C. avellana*, *A. thaliana* and *M. domestica*, produced using MUSCLE with UPGMA clustering.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

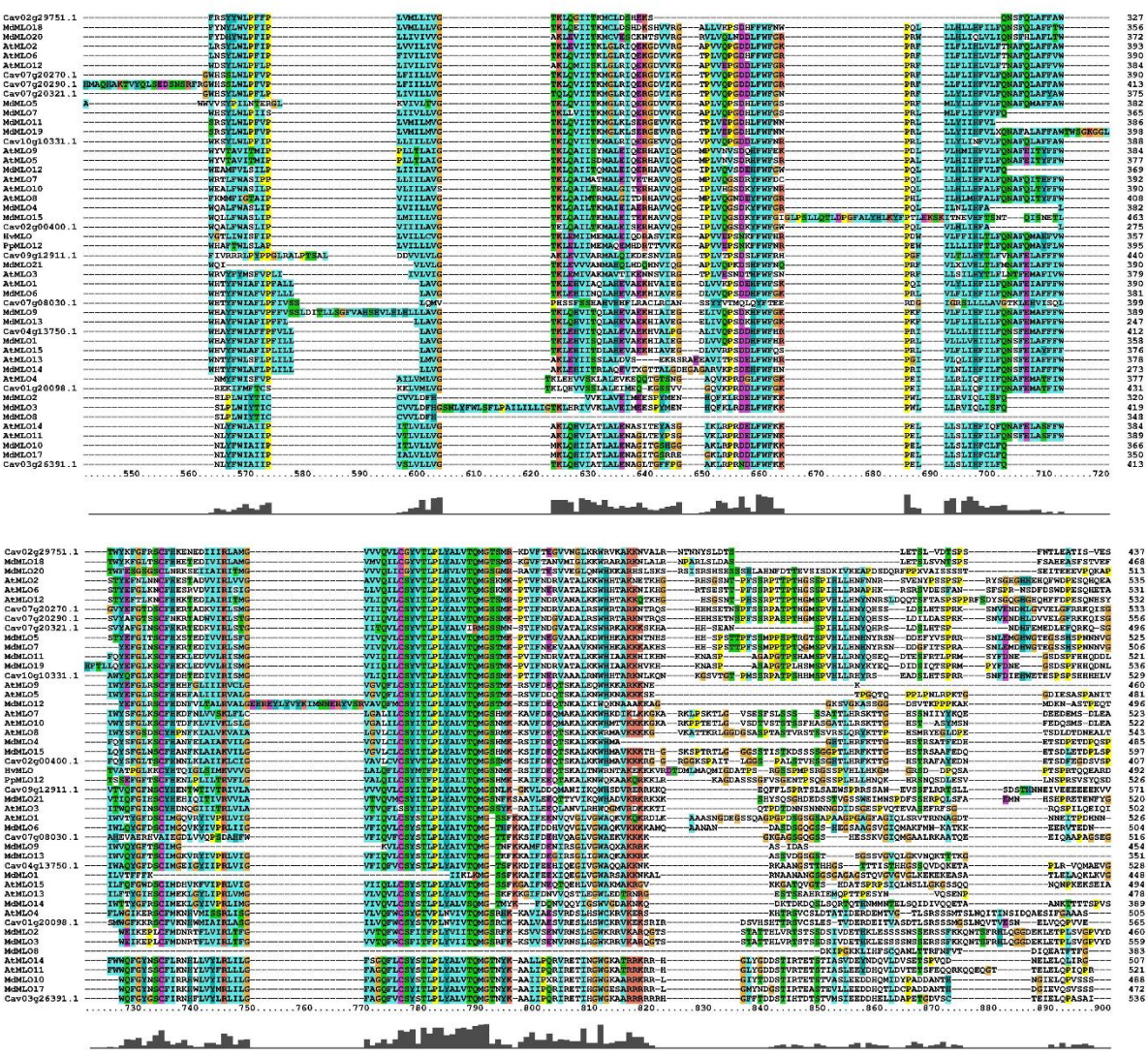


Fig. S6 (continued). Multiple sequence alignment (541-1080 aa) of predicted MLO protein sequences from *C. avellana, A. thaliana* & *M. domestica*. Continued from previous page.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

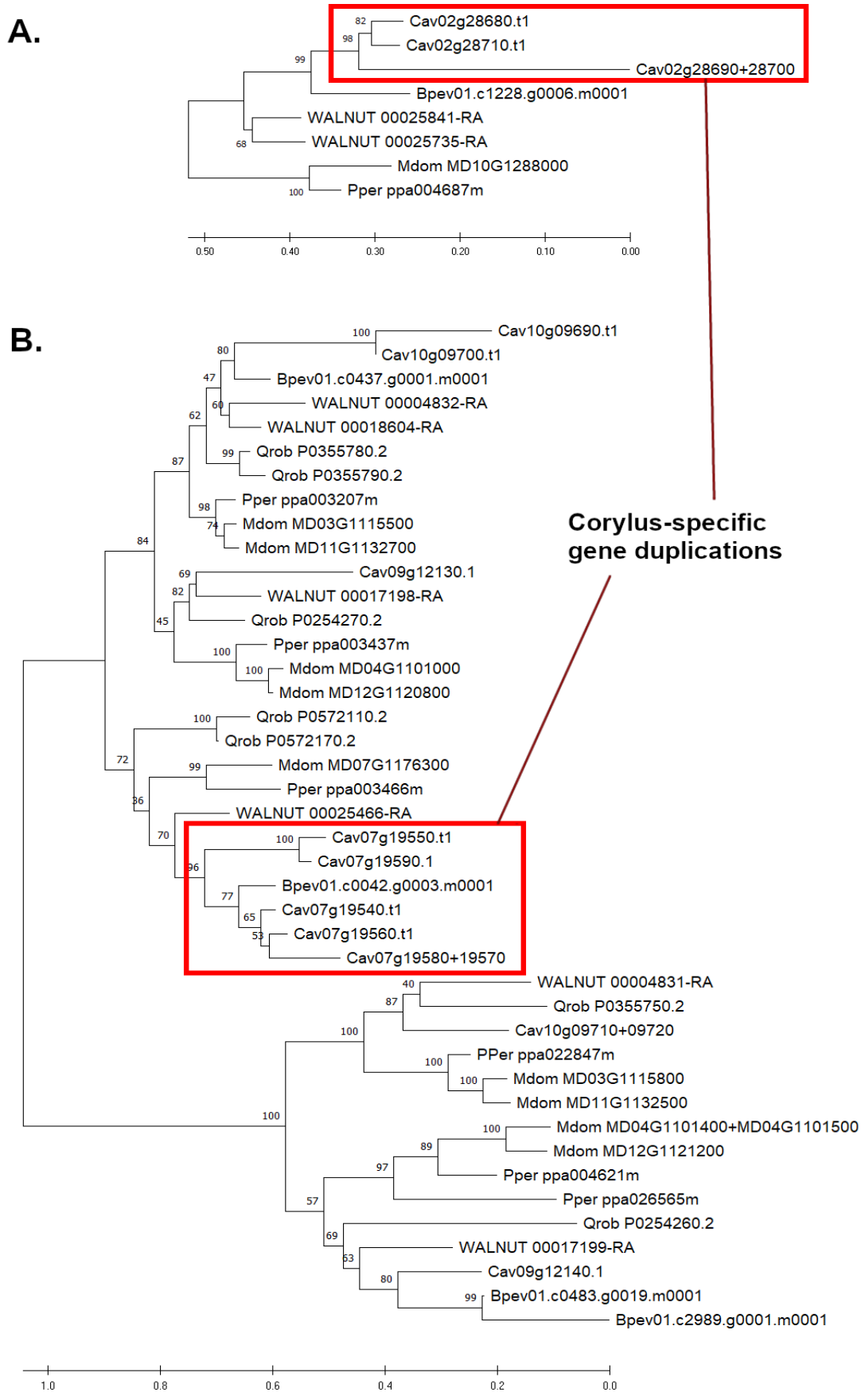


Figure S7. Phylogenetic inference (ML) of tree *MLO*-like genes. **A.** Clade VII, **B.** Clade V + VI

2.11 Identification of Allergen-coding genes and homologs

Gene models encoding most of the reported hazelnut allergens and thier homologs were easily identified from the predicted proteins by sequence similarity. The seed storage proteins (Cor a 9, Cor a 11, Cor a 14) had been identified during *de novo* gene modelling but were not represented in our high-confidence protein set, owing to their absence from our transcriptome data; this is attributed to their exclusive expression in seeds. Complete protein sequences were not available for Cor a 15 or Cor a TLP. For Cor a 15, the genome only contained 4 putative oleosin genes, 2 of which (Cav05g25120 & Cav08g12860) were unambiguously identified as *Cor a 12* and *Cor a 13* respectively. Cav03g05140 encodes a predicted oleosin with a conserved central hydrophobic domain but different N- and C- terminal sequences from Cor a 12, consistent with the description of Cor a 15 [54], whereas Cav06g12950 encodes a putative novel homolog of Cor a 13.

In the case of Cor a TLP, the only sequence information available was that of 2 peptides identified by Edman sequencing[55]. These peptides were both used to search the predicted hazelnut proteins. In total 48 putative TLPs were predicted in the *C. avellana* genome, with a large cluster of 19 highly conserved gene models located in the sub-telomeric region of the short arm of pchr09. One of the peptides matched all of these predicted TLP proteins near their conserved N-terminus with 88% identity; however, the second peptide only gave a partial match (10/17 amino acids) near the C-terminus of some members of the same cluster. Of these the best match was to Cav09g00690, which also encoded the TLP with the highest sequence identity (76.42%) to the known peach TLP allergen Pru p 2.01. Six other genes from the same cluster also matched both Pru p 2.01 and Mal d 2 with >75% identity, and are listed as probable cross-reactive allergens in Table S17.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Supplement References

1. Wang, N. *et al.* Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. *Mol. Ecol.* **22**, 3098–3111 (2013).
2. Andrews, S. & Babraham Bioinformatics. FastQC: A quality control tool for high throughput sequence data. *Manual* (2010) doi:citeulike-article-id:11583827.
3. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu170.
4. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res.* (2009) doi:10.1101/gr.089532.108.
5. Bushnell, B. BBtools. *Joint Genome Institute* (2016) doi:10.1080/00498250701449431.
6. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* (2013) doi:10.1093/bioinformatics/btt476.
7. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, (2019).
8. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* (2015) doi:10.1038/nbt.3122.
9. Colburn, B. C., Mehlenbacher, S. A. & Sathuvalli, V. R. Development and mapping of microsatellite markers from transcriptome sequences of European hazelnut (Corylus avellana L.) and use for germplasm characterization. *Mol. Breed.* **37**, (2017).
10. Kofler, R., Schlötterer, C. & Lelley, T. SciRoKo: A new tool for whole genome microsatellite search and investigation. in *Bioinformatics* vol. 23 1683–1685 (2007).
11. Bassil, N. V, Botta, R. & Mehlenbacher, S. a. Microsatellite Markers in Hazelnut : Isolation , Characterization , and Cross-species Amplification. *J. Am. Soc. Hortic. Sci.* (2005) doi:citeulike-article-id:13527624.
12. Gürcan, K. & Mehlenbacher, S. A. Development of microsatellite marker loci for European hazelnut (Corylus avellana L.) from ISSR fragments. *Mol. Breed.* (2010) doi:10.1007/s11032-010-9464-7.
13. Voorrips, R. E. MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* **93**, 77–78 (2002).
14. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
15. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37**, 7002–13 (2009).
16. Wenke, T. *et al.* Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell* **23**, 3117–28 (2011).
17. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
18. Abouelhoda, M. I., Kurtz, S. & Ohlebusch, E. Replacing suffix trees with enhanced suffix arrays. *J. Discret. Algorithms* **2**, 53–86 (2004).
19. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. in *Methods in molecular biology (Clifton, N.J.)* vol. 1962 1–14 (2019).
20. Falistocco, E. & Marconi, G. Cytogenetic characterization by in situ hybridization techniques and molecular analysis of 5S rRNA genes of the European hazelnut (Corylus avellana). *Genome* **56**, 155–159 (2013).

21. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
22. Lucas, S. J. & Budak, H. Sorting the wheat from the Chaff: Identifying miRNAs in genomic survey sequences of Triticum aestivum chromosome 1AL. *PLoS One* **7**, (2012).
23. Dai, X., Zhuang, Z. & Zhao, P. X. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* **46**, W49–W54 (2018).
24. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008**, (2008).
25. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
26. Fischer, S. *et al.* Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinforma.* **Chapter 6**, Unit 6.12.1-19 (2011).
27. Salse, J., Abrouk, M., Murat, F., Quraishi, U. M. & Feuillet, C. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.* (2009) doi:10.1093/bib/bbp037.
28. Schwacke, R. *et al.* MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol. Plant* **12**, 879–892 (2019).
29. Van Bel, M. *et al.* TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. *Genome Biol.* **14**, R134 (2013).
30. Plomion, C. *et al.* Oak genome reveals facets of long lifespan. *Nat. Plants* (2018) doi:10.1038/s41477-018-0172-3.
31. Chen, F., Mackey, A. J., Stoeckert, C. J. & Roos, D. S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* (2006) doi:10.1093/nar/gkj123.
32. Salojärvi, J. *et al.* Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat. Genet.* **49**, 904–912 (2017).
33. Martínez-García, P. J. *et al.* The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J.* **87**, 507–532 (2016).
34. Jung, S. *et al.* 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res.* **47**, D1137–D1145 (2019).
35. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
36. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
37. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
38. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* (2007) doi:10.1080/10635150701472164.
39. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
40. dos Reis, M. & Yang, Z. Bayesian molecular clock dating using genome-scale

datasets. in *Methods in Molecular Biology* (2019). doi:10.1007/978-1-4939-9074-0_10.
41. Nylander, J. A. A. catfasta2phyml. *https://github.com/nylander/catfasta2phyml* https://github.com/nylander/catfasta2phyml (2018).
42. Larson-Johnson, K. Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. *New Phytol.* **209**, 418–435 (2016).
43. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 407–453 (2015).
44. Manchester, S. R. & Chen, Z. D. A new genus of Coryloideae (Betulaceae) from the Paleocene of North America. *Int. J. Plant Sci.* (1998) doi:10.1086/297569.
45. Gorinsek, B., Gubensek, F. & Kordis, D. Phylogenomic analysis of chromoviruses. *Cytogenet. Genome Res.* **110**, 543–52 (2005).
46. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
47. Gao, D., Li, Y., Kim, K. Do, Abernathy, B. & Jackson, S. A. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. *Genome Biol.* **17**, 7 (2016).
48. Vassetzky, N. S. & Kramerov, D. A. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.* **41**, D83–D89 (2013).
49. Mahesh, H. B. *et al.* Indica rice genome assembly, annotation and mining of blast disease resistance genes. *BMC Genomics* **17**, 242 (2016).
50. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189 (2016).
51. Krupa, A., Anamika & Srinivasan, N. Genome-wide comparative analyses of domain organisation of repertoires of protein kinases of Arabidopsis thaliana and Oryza sativa. *Gene* **380**, 1–13 (2006).
52. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**, (2006).
53. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
54. Nebbia, S. *et al.* Characterization of allergy to hazelnut oleosins in pediatric patients in Italy. *Allergy* **73**, 0308 (2018).
55. Palacín, A. *et al.* The Involvement of Thaumatin-Like Proteins in Plant Food Cross-Reactivity: A Multicenter Study Using a Specific Protein Microarray. *PLoS One* **7**, (2012).

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Contents: *Page*

Supplementary Methods

1.1	DNA and RNA purification	3
1.2	Sequence data processing and genome assembly	3
1.3	Chromosome structure verification & map comparison	4
1.4	Telomeric and Centromeric Repeats	5
1.5	Discovery and masking of repetitive elements	5
1.6	Prediction of hazelnut tRNA, rRNA and miRNAs	7
1.7	Construction and annotation of gene models	7
1.8	Data for comparative genomics and phylogenetic tree construction	9
1.9	Ortholog identification and sequence comparisons	9
1.10	Species and gene set divergence time estimation	10

Supplementary Results and Discussion

2.1	Genome sequencing and hybrid assembly approach	12
2.2	Transcriptome Assembly	14
2.3	Chromosome structure verification & map comparison	14
2.4	Detection of Repetitive elements in the *C. avellana* genome	18
2.5	Annotation and inter-species comparison of tRNA genes	20
2.6	Ribosomal RNA and protein distribution	21
2.7	MicroRNA annotation and target identification	21
2.8	Protein-coding genes and annotation	23
2.9	Comparative Genome Alignment	24
2.10	Characterization of the MLO gene family	25
2.11	Identification of Allergen-coding genes and homologs	31

Supplement References 32

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

List of Tables *Page*

Table S1. PFAM Protein domains used for LTR identification. 6

Table S2. Ortholog Groups containing transposable element domains in the *C. avellana* gene set. 9

Table S3. Fossil calibrations used for calculation of timetrees. 11

Table S4. Genome sequencing and assembly statistics 12

Table S5. Chromosome pseudomolecule statistics 13

Table S6. Sequencing and assembly statistics for RNA-seq libraries 14

Table S7. Summary of Repetitive Elements found in the *C. avellana* genome 18

Table S16. *MLO*-related gene model predictions and results of re-annotation 26

List of Figures

Figure S1. Comparison of SSR marker order on chromosomes with genetic maps. 15-16

Figure S2. Relative abundance of interspersed repeat families in the *C. avellana* genome. 19

Figure S4. Genome statistics of non-coding RNA elements in *C. avellana.* 21

Figure S5. Dot plot representing whole genome alignment of the *C. avellana* and *B. pendula* chromosome assemblies. 24

Figure S6. Multiple sequence alignment (1-540 aa) of MLO protein sequences 27-28

Figure S7. Phylogenetic inference (ML) of selected tree *MLO*-like genes. 30

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Supplementary Methods

1.1 DNA and RNA purification

Genomic DNA was isolated from hazelnut tissues following the method of Wang et al. (2013)[1] with the following modifications: the incubation time with 2x CTAB buffer was shortened to 1 hr at 65^{o}C, and RNAse A was applied in this step rather than as a separate incubation. After resuspension in high-salt TE buffer (1 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA) DNA was bound to a silica membrane column (NucleoSpin Plant II Kit, Macherey-Nagel, Düren, Germany) and washed with 70% ethanol following the kit manufacturer's instructions, to remove low molecular weight DNA fragments, before being eluted in 60 µl of TE buffer.

The same protocol was modified for RNA isolation. All buffers were newly prepared using nuclease-free water. Fresh buds were picked, flash frozen in liquid nitrogen, and ground using 3 mm steel beads in a Tissue Lyser II (Qiagen, Santa Clara, CA). During the 2 x CTAB incubation, RNAse A was not applied; nucleic acids were subsequently recovered by extraction with 24:1 chloroform:isoamyl alcohol. The RNA-containing aqueous phase from this extraction was precipitated using isopropanol. The pellet was cleaned by resuspension first in high salt TE buffer, then in TE without added salt, with ethanol precipitation after each step. The final total RNA was resuspended in 60 µl TE and residual DNA eliminated using RNAse-free DNAse I. In order to confirm that the RNA was not degraded and of sufficient quality for RNA sequencing (RNA-Seq) analysis, 2 aliquots of each sample were also analyzed by agarose gel electrophoresis.

1.2 Sequence data processing and genome assembly

The quality of the raw reads from Illumina and Nanopore sequencing was assessed using FastQC[2].

Illumina only assembly

For Illumina-only assembly, raw sequence reads were processed with Trimmomatic[3] to remove TruSeq adapters, trim bases with quality score <5 from both ends of the reads, and use a sliding window of 4 bases to cut the reads when average sequence quality across the window dropped below 20. The trimmed sequences were then assembled using ABySS 1.9[4] with a range of k-mer size values; `k=80` was found to empirically to give the most contiguous assembly, which was improved further by enabling scaffolding across large bubbles in the k-mer graph (`POPBUBBLES_OPTIONS=--scaffold b=5000`).

Hybrid assemblies

MinION sequencing adapters were trimmed from the first and last 50 nt of NanoPore reads using bbduk from the BBtools suite[5] with the options k=19, editdistance=3. Hybrid genome assembly of the Illumina and trimmed NanoPore reads was carried out using MaSuRCA 3.2[6], with the average Illumina insert size specified as 790±80 nt. This assembly took approximately 9 days running on 36 CPUs in parallel.

Proximity ligation data was integrated with both the Illumina-only and the Hybrid genome assembly using Dovetail Genomics' HiRise assembly pipeline.

Transcriptome mapping

Raw RNA-seq reads were processed with Trimmomatic as described above. Trimmed reads were aligned to the genome using HISAT2 2.1.0[7], using recommended alignment parameters for distinguishing between genotypic variants. StringTie 1.3.4[8] was employed to assemble the transcripts from RNA-seq reads aligned to the genome, and compute their abundance.

1.3 Chromosome structure verification & map comparison

The SSR-based genetic map for OSU 252.146 x OSU 414.062 [9] was compared with the chromosomes as follows: firstly, all SSRs within the chromosome sequences were identified, and each locus extracted along with 500 nt of upstream and downstream sequence using SciRoKo 3.2[10]. Next, PCR primer sequences for 290 previously mapped SSR markers were

collated[9,11,12]. The SSR flanks were searched for these primer sequences using blastn (-`task blastn-short`) and requiring >95% sequence identity. When both primers matched the flanks of the same SSR locus, this was taken as a positive match, and the chromosome position of the locus compared with the genetic map position of the same SSR (Supplementary Figure 1). Genetic map/chromosome comparisons were drawn using MapChart 3.2[13].

1.4 Telomeric and Centromeric repeats

Chromosome pseudomolecules were searched for classical plant telomere repeats using BLASTN with $(GGGATTT)_5$ as the query seqeunce. In order to identify putative centromeric regions, the chromosomes were searched using tblastn for integrase domain protein sequences specific to centromere-associated CRM class LTRs from *Zea mays* and *Hordeum cereba* (Genbank Accession Nos: AY129008 & AY040832). Matches with < 60% sequence identity or < 80% query coverage were eliminated.

1.5 Discovery and masking of repetitive elements

For identification of candidate novel retrotransposons, the following repeat-type specific prediction tools were used: LTRharvest/LTRdigest for LTRs and TRIMs[14,15], SINE-finder for SINEs[16], TRF for arrays of short tandem repeats[17]. In order to identify novel LTR elements, Hidden Markov Models (HMMs) of known LTR proteins were downloaded from Pfam (Table S1) and supplied as a search library for LTRdigest. Novel LTR elements were clustered using Vmatch[18] following the 80-80-80 principle (80% sequence identity with over 80% coverage of the shorter element, and > 80 bp). After selecting one representative of each cluster, degenerate copies were identified using blastn searches separately for the 5' terminal repeat, 3' terminal repeat, and the internal region of each element. Those with at least 5 LTR copies, or at least 3 copies of the internal region, were accepted as novel LTRs.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Table S1. PFAM Protein domains used for LTR identification.

Pfam ID	Pfam name	Description
PF07253	Gypsy	Gypsy protein
PF00077	RVP	Retroviral aspartyl protease
PF08284	RVP_2	Retroviral aspartyl protease
PF00078	RVT_1	Reverse transcriptase
PF07727	RVT_2	Reverse transcriptase
PF06817	RVT_thumb	Reverse transcriptase thumb domain
PF06815	RVT_connect	Reverse transcriptase connection
PF00075	Rnase H	Ribonuclease H domain
PF02022	Integrase_Zn	Integrase Zinc binding domain
PF00665	rve	Integrase core domain
PF00098	zf-CCHC	Zinc knuckle domain
PF00385	Chromo	'Chromo' (CHRomatin Organisation MOdifier) domain
PF01393	Chromo_shadow	Chromo shadow domain
PF00692	dUTPase	dUTPase domain
PF03078	ATHILA	ATHILA ORF-1 family
PF04195	Transposase_28	Putative Gypsy type transposon
PF05380	Peptidase_A17	Pao retrotransposon peptidase
PF01577	Peptidase_S30	Potyvirus petidase
PF13975	gag-asp_protease	Retroviral aspartyl protease
PF13976	gag_pre-integrs	Retroviral Gag
PF03732	Retrotrans_gag	LTR gag (Gypsy type)
PF14223	Retroran_gag_2	LTR gag (Copia type)
PF18103	SH3_11	Retroviral Int
PF00552	IN_DBD_C	Retroviral Int

These newly identified repeats were combined with known eudicot repetitive elements from RepBase Update 22.08 (5,913 elements) and mipsREdat 9.3 (26,123 elements) to create a custom repeat library. This was then used to search the genome by RepeatMasker 4.0.7 on 'normal' sensitivity with default scoring matrices. Detected repeats were masked with runs of 'N'. For subsequent identification of protein-coding genes and other functional elements, the `-nolow` option was used to leave simple repeats and low-complexity regions unmasked.

1.6 Prediction of hazelnut tRNA, rRNA and miRNAs

Genes encoding tRNAs were identified using tRNAscan-SE 2.0.0[19].

To find rRNA genes, the masked chromosome sequences were searched using BLASTN (e-value cut-off 1e-30) with the following coding sequences: previously published 5S rRNA of *C. avellana* (Genbank HF542974.1[20]); and complete tree 45S rRNA sequences retrieved from Genbank. The pseudochromosomes were searched for sequences with similarity to known ribosomal proteins from other plants recorded in NCBI (listed in Table S9) using BLASTN. Genome regions passing a match threshold (>80 % identity, >80 % query coverage and >200 bp in length) were selected, and putative protein sequences searched using BLASTX (e-value cutoff 1e-6) against the NCBI non-redundant plant protein database. Sequences with a closest match to non-ribosomal proteins were removed.

Conserved miRNA genes were identified using all reported plant miRNAs from miRBase v22 [21] with SUmirFind and SUmirFold[22], using a mismatch cutoff ≤ 3 for initial miRNA homolog detection, followed by predicted pre-miRNA secondary structure prediction and selection of strong miRNA candidates based on established structural criteria[22]. For inter-species comparison, the same methods were used to identify miRNA genes in other plant genomes. Expression of putative miRNAs was confirmed by using all non-redundant pre-miRNA sequences to search assembled Tombul transcripts using BLASTN, retaining hits with >78% identity, >80% coverage as expressed pre-miRNAs. Experimentally validated targets of specific miRNA families found in *C. avellana* were retrieved from miRBase. Predicted targets of putative miRNAs were found in the transcripts using psRNATarget[23], and annotated with Gene Ontology terms using Blast2GO software[24].

1.7 Construction and annotation of gene models

Augustus software was used to generate *ab initio* gene predictions from the masked Tombul genome based on Hidden Markov Models optimised for *Arabidopsis thaliana* [25].

Evidence for gene expression was produced by aligning the gene models to the Tombul transcriptome using BLASTN, retaining only the best hit for each gene model. Conserved orthologs of known plant genes were identified using the 'Assign proteins to groups' pipeline of OrthoMCL-DB (https://orthomcl.org/orthomcl/)[26]. Multiple BLAST hits between different segments of the same pair of sequences were merged using a custom Perl script to calculate the 'Cumulative Identity Percentage' (CIP) and 'Cumulative Alignment Length Percentage' (CALP)[27]. Low-quality hits were eliminated based on the following criteria: CIP < 90%, CALP <60% for transcript matches, and CIP < 70%, CALP < 60% for ortholog matches. Gene models were selected as 'high-confidence' if they were supported by a good quality match to a transcript sequence and/or to an orholog from other plant species. The OrthoMCL analysis also identified 12,516 genes that formed clusters containing only gene models from *C. avellana,* with no orthologs from the other plant species in OrthoMCL-DB. These putative hazelnut-specific genes were added to the high-confidence list if: i) they had a transcript match at CIP >90% and CALP >90%, showing that they are expressed; or ii) they matched an expressed high-confidence gene at CIP > 70% and CALP > 60%, suggesting that they could be paralogs resulting from gene duplication. This gave a working set of 27403 gene models. Automated functional annotation of the gene models was carried out using Mercator4, and TRAPID2 with default parameters[28,29]. Gene models annotated as retroviral/retrotransposon proteins (Reverse Transcriptase, retrotransposon Gag, chromodomain) or assigned to OGs corresponding to these proteins were investigated (Table S1), and gene models that lacked transcript support were eliminated as probable degenerate repeats[30]. The final edited gene set contained 27270 gene models.

Table S2. Ortholog Groups containing transposable element domains in the *C. avellana* gene set.

Group ID from OrthoMCL-DB[31]	Domain description	No. of orthologs in *C. avellana* gene set	No. of orthologs with transcript evidence
OG5_126614	Reverse transcriptase	3	0
OG5_131059	MULE transposase	8	5
OG5_131652	Reverse transcriptase	1	0
OG5_138969	Retrotransposon gag protein	3	0
OG5_126614	Reverse transcriptase	6	0
OG5_153395:	Plant Mobile Domain	2	2
OG5_156509:	Plant transposase (Ptta/EnSpm)	0	0
OG5_159949:	Reverse transcriptase	2	0
OG5_184563:	Plant transposase (Ptta/EnSpm)	0	0
OG5_212464:	Plant transposase (Ptta/EnSpm)	0	0

1.8 Data for comparative genomics and phylogenetic tree construction

The final set of high-confidence gene models from *C. avellana* var. 'Tombul' was compared with other trees from the order Fagales for which complete genome assemblies and annotated gene sets are available: silver birch (*Betula pendula*)[32], English walnut (*Juglans regia* cv. 'Chandler')[33], and pedunculate oak (*Quercus robur*)[30]. As an outgroup, two well-characterized fruit tree genomes, apple (*Malus domestica* cv. 'Golden Delicious') and peach (*Prunus persica*) were selected from the Rosaceae[34]. Whole genome sequence comparisons were carried out using LAST v869[35].

1.9 Ortholog identification and sequence comparisons

The complete sets of protein sequences for all annotated genes in these six genomes were downloaded and compared with each other using OrthoFinder 2.3.11, using the default pipeline to find orthogroups and infer rooted gene trees for each group[36]. The automatically generated species tree incorrectly assigned *Q. robur* as the outgroup; therefore the tree was modified to place the root on the branch containing the divergence between *Q. robur* (and

the other Fagales) and the Rosales. The final step of the OrthoFinder analysis, reconciling gene trees with the species tree, was then re-run using the correctly rooted tree.

For genome-wide inference of divergence times, the set of 979 single-copy orthologs (SCOs; genes found as a single copy in all 6 species) was selected. Separate fasta files were produced by OrthoFinder containing the 6 predicted protein sequences for each SCO; seqeunces within each file were aligned with each other using ClustalW[37]. Sequence segments that are only present in 1 or 2 orthologs can skew phylogenetic analysis, including timetrees, so alignments were edited using Gblocks 0.91[38], retaining only sequence blocks that were conserved in at least 4/6 species. 113 SCOs for which <30% of the total sequence length was in conserved blocks were considered as suspect/low quality and were removed from the analysis, leaving 866 edited SCOs for the following steps.

1.10 Species and gene set divergence time estimation

Time-scaled phylogenetic trees were constructed using the MCMCTree program from the PAML 4.9j package (Phylogenetic Analysis by Maximum Likelihood)[39]. MCMCTree carries out Bayesian inference of divergence times on a sequence alignment using a variety of relaxed molecular clock models, and a guide phylogenetic tree including soft fossil calibrations[40]. The edited gene sequence alignments created above were concatenated and converted into sequential Phylip format as input for MCMCTree using catfasta2phyml.pl[41]. Calibration points for the guide tree were selected from previous detailed surveys of fossils from the Fagales[42] and all angiosperms[43] and are summarized in Table S3. For each divergence of the tree, the most recent possible age of the oldest extant fossil(s) within daughter lineages were used to set a hard minimum age for the crown group. The extinct species *Cranea wyomingiensis* was originally described as belonging to the Coryloideae[44]; however it also has morphological characters of other Betulaceae genera, so here was left in the Betulaceae crown group.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Table S3. Fossil calibrations used for calculation of timetrees. The 4th and 5th columns indicate the branch points to which calibrations were added, and the figures used.

Fossil species	Age estimate (mya)	Taxon	Divergence point on timetree	Minimum age (mya)
Betula leopoldae *Alnus* sp.	37.2 - 48.6 48.6 - 55.8	Betuloideae	*Betula* \ *Corylus*	49
Carpinus perryae *Corylus johnsonii*	49 - 50 49 - 50	Coryloideae		
Cranea wyomingiensis	55.8 - 65.5	Betulaceae	*Betula & Corylus* \ *Juglans*	55.8
Cyclocarya brownii *Polyptera manningii*	55.8 - 65.5 56.8 - 66.3	Juglandaceae		
Prunus wutuensis *Oemleria janhartfordae*	48.6 - 55.8 49.4 - 54.2	Rosaceae	*Malus* \ *Prunus*	49
Archaefagaceae futabensis	87.5 - 96.3	Fagales	Rosales \ Fagales	87.5

The MCMCTree analysis was realized using a relaxed clock model with independent rates on each branch of the tree following a log-normal distribution, and the root age constrained to <120 million years ago (mya). The JC69 substitution matrix was used and the fossil ages indicated in Table S3 were set as soft minimum bounds for each node using the default Cauchy probability distribution (see PAML documentation, except that the scale factor was reduced from 1 to 0.5. The Markov chain was allowed to burn-in for 4000 iterations, and samples were then collected every 20 iterations until a total of 20,000 had been collected. The MCMCTree process was run twice and the dating estimates from both runs compared; stable time and rate estimates during the run, and highly consistent values between runs (R^2 > 0.995) were taken to indicate that the Markov chains had converged.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Supplementary Results and Discussion

2.1 Genome sequencing and hybrid assembly approach

As discussed in the main text, 3 different whole-genome shotgun sequencing technologies were used to sequence the *C. avellana* var. 'Tombul' genome; short paired-end reads (Illumina PE), long reads (NanoPore) and proximity ligation sequencing (Dovetail Chicago-HiC). Genome assembly was carried out for Illumina only and for each combination of 2 or 3 of the technologies (Table S4). The 'Illumina only' assembly was highly fragmented and ~25% larger than the expected genome size (378 Mb), probably due to heterozygous regions being assembled twice into separate contigs. A hybrid assembly of the Illumina and NanoPore data (Illumina + NP) gave 12,557 scaffolds that were ~10-fold larger than 'Illumina only' across the size distribution and eliminated the large majority of duplicated sequences and gaps in the assembled scaffolds. However, this assembly was still too fragmented (N90 = 5857) to allow large-scale structural comparisons. In both genome assemblies, the observed GC content was 36%.

Table S4. Genome sequencing and assembly statistics

Sequencing technology	**No. of Reads**	**Library design**	**Total read length (Gb)**	**Sequence coverage[1*]**
Illumina PE	2 x 136M	Paired end, 700-800 bp insert	41.1	108 x
NanoPore	1,351,274	Single molecule, 1-10 kb reads	3.53	9.3 x
			Insert size range	Physical coverage[2*]
Dovetail Chicago	2 x 221M	Proximity ligated	1-100 kb	225 x
Dovetail HiC	2 x 242M	Proximity ligated	10 kb – 10 Mb	3,447 x

Assembly Statistics				
	Illumina only	Illumina + Dovetail	Illumina + NP	Illumina + NP + Dovetail
No. of scaffolds (>1kb)	89,427	32,741	12,557	2,206
Scaffold N50 / L50	20927 / 7.32 kb	12 / 13.33 Mb	1299 / 78.8 kb	5 / 36.65 Mb
Scaffold N90 / L90	149112 / 204 bp	89585 / 3.4 kb	5857 / 13.7 kb	10 / 22.72 Mb
Total scaffold size (Mb)	513.8	520.1	383.1	384.2
% gap bases (N)	6.01%	6.45%	0.18%	0.47%
Largest scaffold	0.152 Mb	45.9 Mb	0.732 Mb	50.95 Mb

[1]Average no. of times a genome nucleotide is included in a sequence read

[2]Average no. of tımes a genome position is included in the region between 2 linked reads

* Calculated for an estimated genome size of 378 Mbp.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

In order to improve the contiguity of the assembly, we carried out proximity ligation sequencing using Dovetail Genomics' proprietary methods. These generate pairs of linked reads that originate from within the same large DNA fragment (Chicago library) or from physically adjacent nucleosomes in native chromatin (HiC library). The sequences of the linked reads are not assembled directly, but mapped to the scaffolds from the pre-existing genome assembly. The 'HiRise' bioinformatic pipeline then uses these links to determine the order and orientation of scaffolds along each pseudomolecule. Adjacent, non-overlapping scaffolds are joined with an arbitrary gap sequence of $(N)_{100}$. As shown in Table 1, incorporation of Dovetail data enabled >50% of the genome to be assembled into chromosome-sized pseudomolecules both from Illumina-only and Illumina + NP assemblies. However, a very large number of small scaffolds remained unassembled in the 'Illumina + Dovetail' assembly, reflected by the very high N90 (Table S4), and the duplicate scaffolds were not resolved. In contrast, the Illumina + NP + Dovetail assembly consisted of 11 chromosome-sized pseudomolecules ranging from 22.42 – 50.95 Mb in length (Table S5), in total accounting for 97.8% of the predicted genome size. The next largest scaffold after these contained the chloroplast and mitochondrial DNA, after which there remained 2194 unplaced small scaffolds (1 kb - 400 kb) with a cumulative length of 13.6 Mb.

Table S5. Chromosome pseudomolecule statistics

Pseudo-molecule ID	Size (Mbp)	Size group	No. of gaps ('N' percentage)	OSU linkage group (Size in cM)	Matched SSRs (no. / no. colinear)
pchr01	50.95	L	1337 (0.48%)	1 (193)	16 / 16
pchr02	50.86	L	1565 (0.54%)	2 (185)	10 / 10
pchr03	39.77	M	1010 (0.43%)	4 (153)	14 / 12
pchr04	36.85	M	914 (0.47%)	9 (184)	14 / 13
pchr05	36.65	M	1031 (0.46%)	5 (123)	11 / 11
pchr06	30.27	M	816 (0.47%)	7 (123)	9 / 8
pchr07	30.24	M	609 (0.49%)	11 (118)	6 / 6
pchr08	25.77	S	665 (0.46%)	3 (105)	10 / 10
pchr09	23.27	S	465 (0.38%)	10 (101)	8 / 8
pchr10	22.72	S	557 (0.55%)	6 (88)	2 / 2
pchr11	22.42	S	630 (0.49%)	8 (103)	13 / 12
Total	369.78		9599 (0.48%)		113 / 108

2.2 Transcriptome Assembly

Two RNA-seq libraries were generated from total RNA isolated from leaf buds, and were sequenced by following Illumina protocols with results as indicated in Table S5. After quality control and trimming, the reads were aligned to the chromosome pseudomolecules using Hisat2. Aligned reads were assembled using StringTie individually for each sample, and then redundant transcript contigs from both samples were combined, using StringTie's 'merge' function; at this step, putative transcript isoforms with very low expression levels (<1% of all expression for that transcript) The final assembly included 57315 distinct transcripts arising from 47250 transcribed 'genes' (Table S6). This is expected to be greater than the number of protein-coding genes in the genome, because it includes non-coding RNAs; also, heterozygous alleles may be assembled twice by this method, into different 'genes'.

Table S6. Sequencing and assembly statistics for RNA-seq libraries (Paired end, 2 x 100 nt)

Sample ID	**Total read bases (bp)**	**Total reads**	**GC(%)**	**AT(%)**	**Q20(%)**	**Q30(%)**
Library_T1	6,677,685,902	66,115,702	46.310	53.69	98.763	95.400
Library_T2	7,017,024,288	69,475,488	46.084	53.92	98.163	95.286
Mapping & assembly	Filtered reads	Reads mapped & paired	% Properly paired	Assembled transcript contigs	Merged isoforms	Final 'genes'
Library_T1	65,937,538	61,974,596	94.0	76540	57315	47250
Library_T2	69,238,960	62,411,216	90.1	77602		

2.3 Chromosome structure verification & map comparison

The chromosome arm structure of the Tombul pseudomolecules was verified by comparison with existing genetic maps (Supplementary Figure 1) and identification of telomeric and centromeric repeats. Telomeres are difficult to assemble due to their highly repetitive nature; indeed classical plant telomere repeats $(GGGATTT)_n$ were only observed at the ends of 4 of the chromosome pseudomolecules, while a possible non-canonical telomere sequence (GGGGTTTCATTT) was found at the ends of 2 further pseudomolecules.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

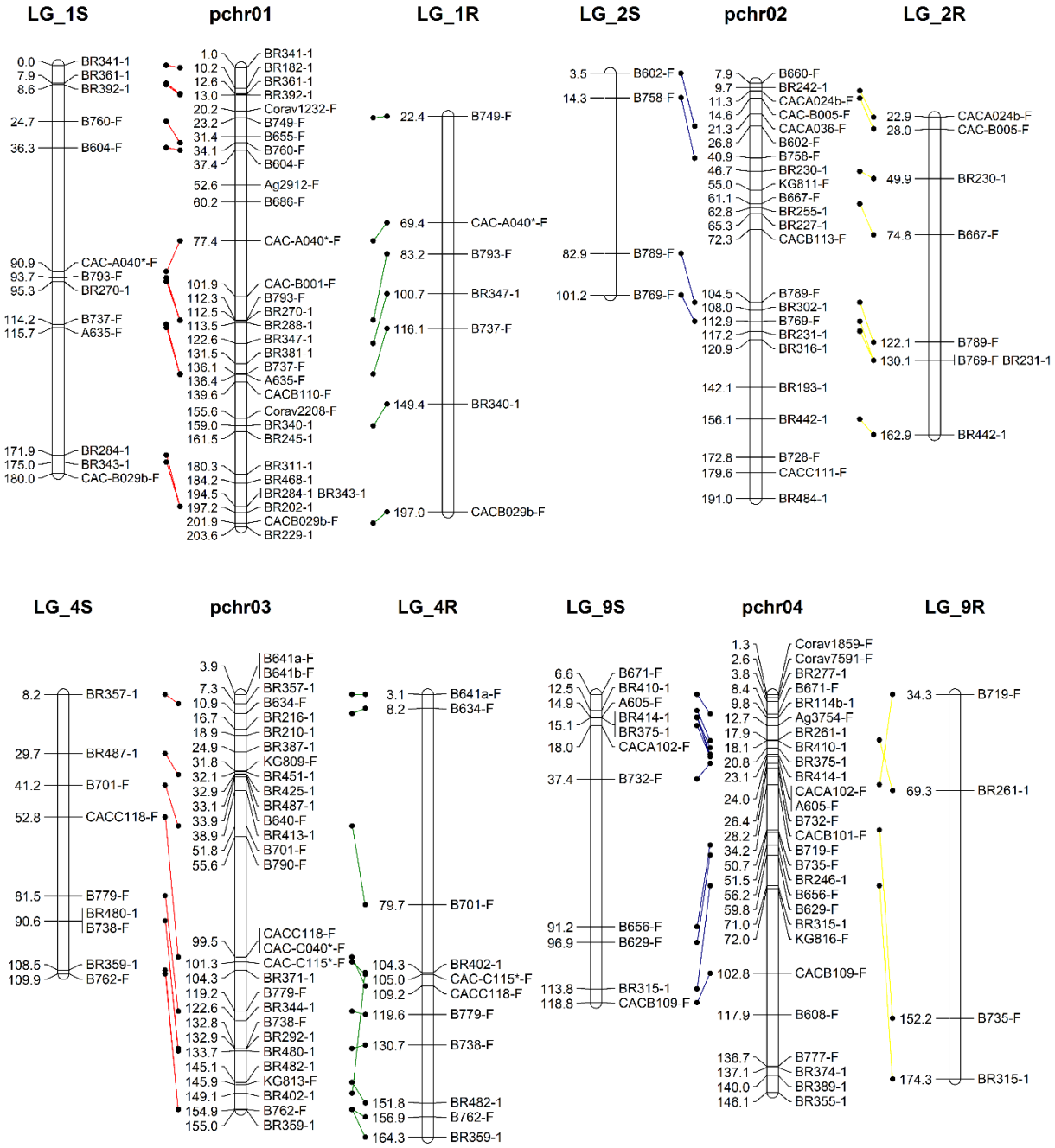
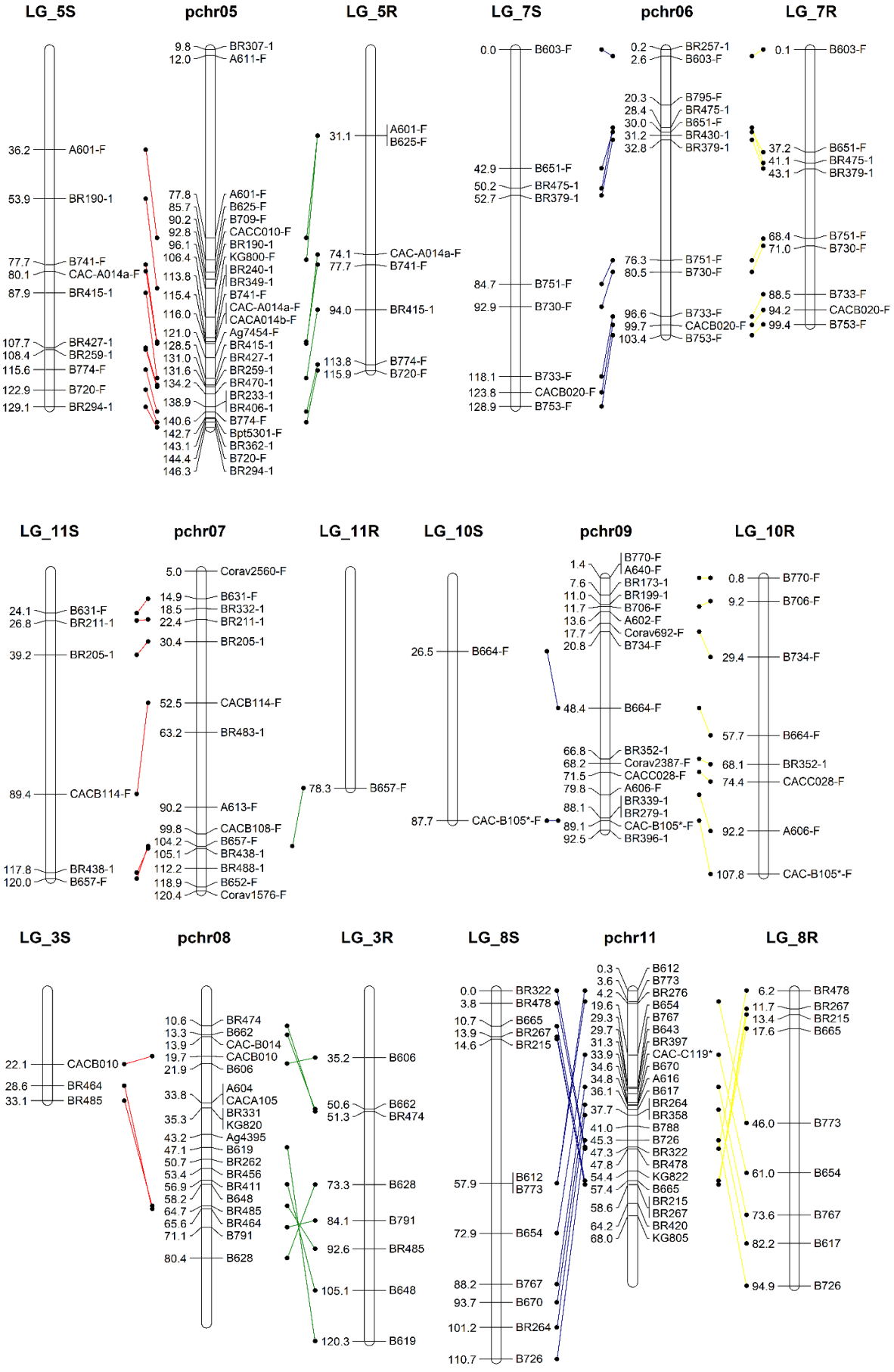


Figure S1. Comparison of SSR marker order on chromosomes with genetic maps. LG_1S/1R indicate the linkage groups from the OSU 252.146 x OSU 414.062 cross, with marker positions marked in cM, and only showing markers also found on the chromosomes. To give a similar scale, the markers on the chromosome bars are labelled with their bp position/250,000. The orientation of markers on mis-assembled chromosomes (next page) showed the necessary correction; in pchr08, the chromosome arms are inverted; in pchr11, they are correctly orientated but the long arm is stacked above the short arm. The map for LG_6S/6R & pchr10 is not shown owing to the small number (2) of shared markers.

Supporting Information

“A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement” Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)



The CRM sub-family of Gypsy LTR retrotransposons has a uniquely conserved integrase domain (the 'chromodomain') and is associated with peri-centromeric regions in plants[45]. Searching for this domain found 2-3 copies near the centre of 8 of the chromosomes, while no high-confidence match was found on pchr02. Notably, the 2 remaining pseudomolecules (pchr08 & pchr11) appeared to have CRM family repeats at both ends and telomeric repeats near the centre, suggesting that the chromosome arms were erroneously joined at the telomeres rather than the centromere.

Comparison with the published genetic map (Supplementary Figure 1, bottom) confirmed this; in these two pseudomolecules, the SSRs mapped to the halves of each chromosome were co-linear with those in the corresponding linkage group, but the order of the 2 arms was reversed. In both cases the false join was determined to have been made at the final assembly step (Dovetail HiC) where two scaffolds that both had classical telomere repeats at one end were joined together. Accordingly, these two chromosome sequences were manually split at the telomeric repeats, the halves inverted and re-joined according to the order of SSR markers shown by the genetic map.

This assembly error can be explained as an artifact of the HiC ligation method; if the two ends of the chromosome are located in close physical proximity in native chromatin, links could be created between the two sub-telomeric regions. Similarly, in native chromatin few links may be produced between the highly condensed peri-centromeric regions, even though they are adjacent. In these circumstances, the HiRise pipeline could mistakenly join the chromosome arms at the telomeres rather than the centromere.

After rearrangement, 108 of 113 SSRs shared between the chromosomes and the OSU genetic linkage map were co-linear.

2.4 Detection and Classification of Repetitive elements in the *C. avellana* genome

About a third of the Tombul genome was found to consist of LTR retrotransposons (summarized in Table S7), the majority of which were copies of 598 novel LTR elements identified in this study (Supplementary Methods). New Gypsy & Copia superfamily elements were distinguished by the order of protein-coding domains in their internal regions[46]. When insufficient protein-coding domains were present positively to identify the family, the elements were classified simply as 'LTR'.

Table S7. Summary of Repetitive Elements found in the *C. avellana* genome

	Number of elements	**Length occupied (bp)**	**Percentage of sequence (%)**
Retroelements	350,766	131,114,230	35.38
SINEs:	239	14,731	0
LINEs:	620	103,692	0.03
- RTE/Bov-B	96	19,947	0.01
- L1/CIN4	551	99,487	0.03
LTR elements:	349,907	130,995,807	35.35
- Ty1/Copia	101,052	32,516,429	8.78
- Ty3/Gypsy/DIRS1	50,068	25,164,581	6.79
Retroviral	0	0	0
DNA transposons	2,397	1,123,336	0.3
hobo-Activator	787	354,147	0.1
En-Spm/CACTA	385	164,418	0.04
MuDR-IS905	438	182,472	0.05
Tourist/Harbinger	234	98,731	0.03
Helitron	488	269,216	0.07
Unclassified:	589	106,975	0.03
Total interspersed repeats:		**132,344,541**	**35.72**
Small RNA:	399	41,681	0.01
Satellites:	7	579	0
Other simple repeats:	208820	7,134,639	1.93
Low complexity:	36677	1,756,537	0.47
Total simple repeats:		**8,933,436**	**2.41**
Total repetitive content:		**141,262,718**	**38.13**

TRIMs were identified using the following criteria[47]: i) having terminal repeats and 4-6 bp target site duplications, simlar to LTRs, ii) maximum length 1500 bp, iii) containing no retrotransposon protein coding genes, and iv) having at least 10 copies in the genome. Putative novel SINEs were required to have >60% sequence identity over at least 60 bp to a known tRNA gene, and multiple full-length copies in the genome[48].

In total the chromosome pseudomolecules were found to consist of 35.72% interspersed repeats, and 2.41% simple repeats (Table S6, Figure S2).

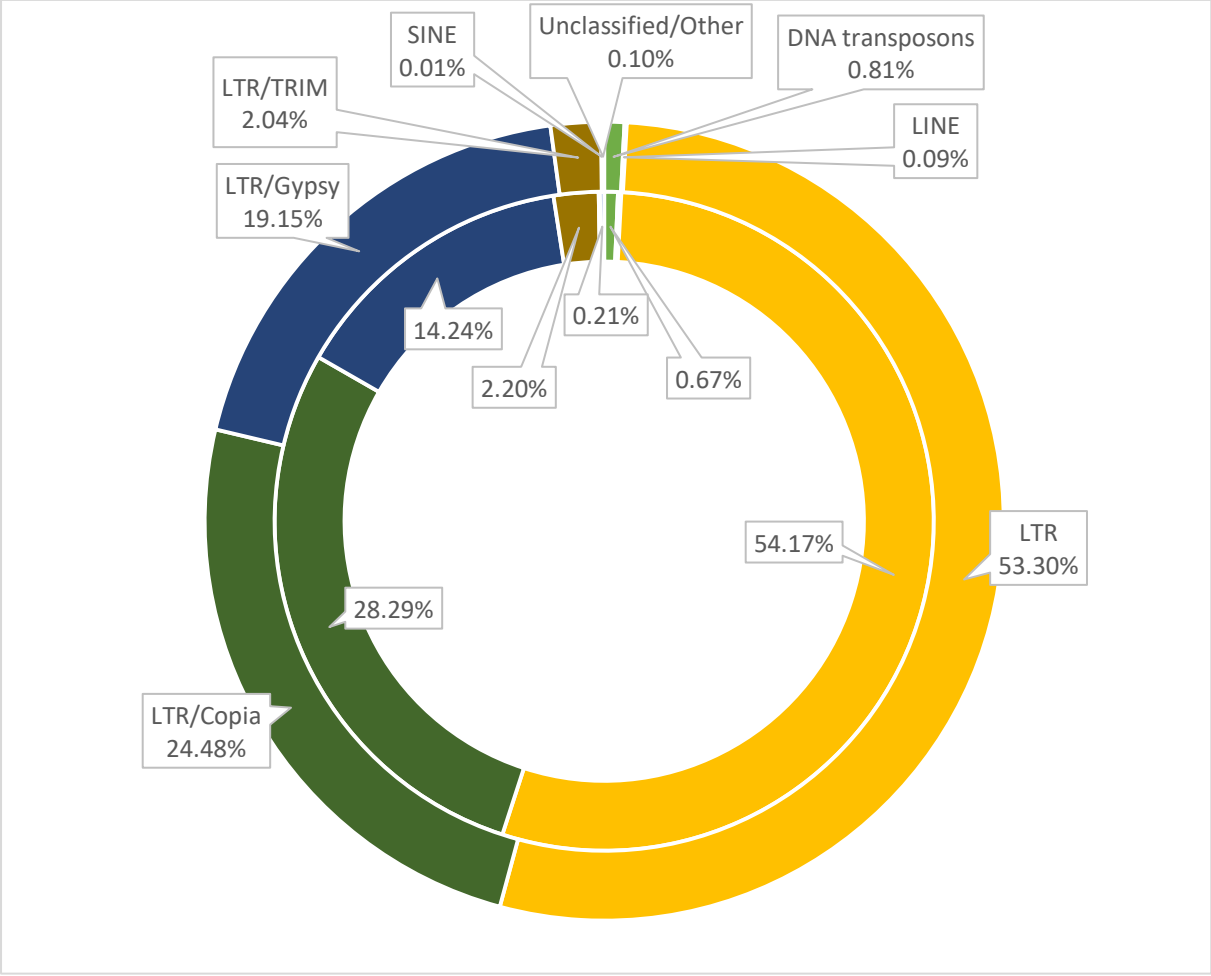


Figure S2. Relative abundance of interspersed repeat families in the *C. avellana* genome. Outer ring shows percentage of total sequence coverage by interspersed repeats (132.3 Mb from the whole genome); inner ring shows percentage of total number of interspersed repeat loci. LINEs and SINEs are not indicated on the inner ring due to their low abundance.

The overall abundance of repetitive elements is similar to the rice genome (35-40% depending on genome assembly[49]) but lower than that of the slightly larger silver birch genome (*Betula pendula* L., 51.35%[32]). Over half of the LTR elements were too diversified to be positively assigned to a sub-family; of the rest, Copia elements were almost twice as abundant as Gypsy elements, but owing to their shorter average length, the total number of bases covered by both families was comparable. TRIMs (Terminal Repeats In Miniature) are a relatively recently described class of short, non-autonomous LTR retroelements, that are thought to replicate using the enzymes encoded by related Copia or Gypsy LTRs[47]. TRIMs were found to be highly abundant in the *B. pendula* genome with over 82,000 copies, but their copy number in *C. avellana* was only 8,712, which is within the typical range observed for other plant genomes. Terminal inverted repeat (TIR) DNA transposons, which were again abundant in *B. pendula,* were notably scarce in *C. avellana*; consistent with this, there were no intact copies of the EnSpm/Ptta transposase, responsible for replicating the TIR/CACTA transposon family, detected in the genome (Table S2). These observations suggest that the expansion of these repeat families in *Betula* took place after the ancestors of the *Betula* and *Corylus* lineages diverged.

2.5 Annotation and inter-species comparison of tRNA genes

A total of 477 predicted tRNA genes and 40 tRNA pseudogenes, with an average length of 75 nt, are listed in Table S7. Of these 27 tRNA genes were predicted to contain introns, chiefly some of those encoding Gly, Met & Tyr. Codon frequencies for tRNA genes in *C.avellana* were calculated and compared with other tree species and Arabidopsis (Table S8). tRNA predictions for *A. thaliana* were taken from GtRNAdb[50].

Various differences in abundance between tRNA genes for particular amino acids, and codons for the same amino acid, were observed (Supplementary Figure 3); for example, tRNAs for codons ending "–AT" were non-existent or rare in all the plant genomes examined. Some species-specific differences were also noted; *C.avellana* has a higher abundance of tRNAs for

all Glycine codons, and especially GGC, than the other plant genomes, whereas *P. trichocarpa* has an increased number of all Alanine & Leucine tRNAs. Among the Serine codons, *C. avellana* has a greater abundance of tRNAs for AGC than the other tree species, whereas Arabidopsis favours TCT. Further study would be useful to determine whether these differences are correlated to codon frequencies in protein-coding genes, or influence translation efficiency in different plant species; however, overall there were few species- or lineage-specific variations in tRNA abundance.

2.6 Ribosomal RNA and protein distribution

The only full-length 45S rDNA was identified on pchr08, however, there was also a partial rDNA locus on pchr02. 21 copies of the 5S rDNA were identified on 7 different chromosomes, with the largest cluster of 8 tandemly repeated copies on the short arm of pchr05, followed by a cluster of 3 copies on pchr11 (Table S9). Homologs of ribosomal proteins from other plant species were also found distributed across most of the chromosomes, but absent from pchr11 (Figure S4A & Table S9). In contrast, tRNA genes were relatively evenly spread among the chromosomes, while miRNA families were largely absent on pchr09, suggesting that these different classes of non-coding RNA elements have undergone independent assortment and expansion during the evolution of the *C. avellana* genome.

2.7 MicroRNA annotation and target identification

153 candidate conserved miRNA genes were predicted in the *C. avellana* genome based on previously established annotation criteria (Table S7). The copy number of conserved miRNA families was determined by removing redundant matches from multiple query miRNAs in the output of SUmirFold, to identifying unique miRNA-coding sequences. In some cases, several mature miRNAs from other plants matched the same pre-miRNA candidate; for example, a miR171 family gene located from 944161-944280 on pchr01 had 3 possible mature miRNA predictions from different parts of the pre-miRNA. In these cases all candidates were retained

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

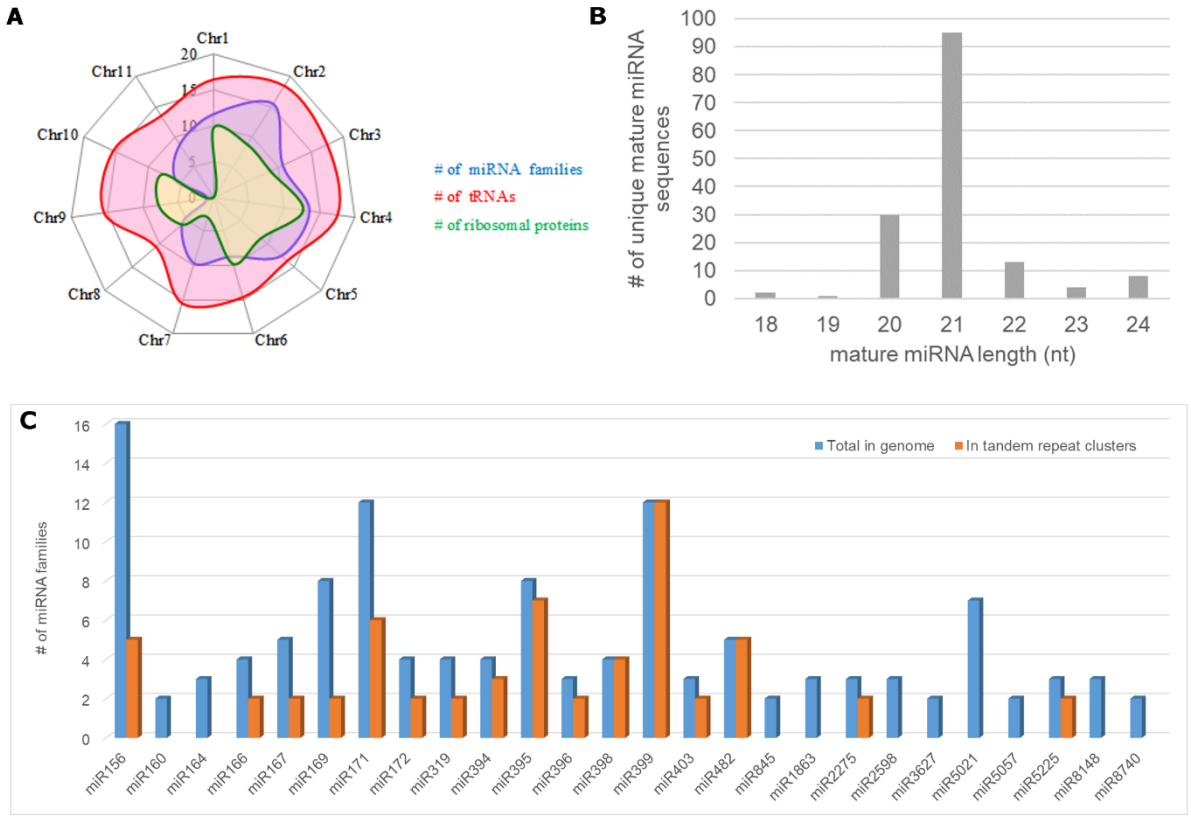


Figure S4. Genome statistics of non-coding RNA elements in *C. avellana.* **A.** Radar plot of the distribution of ribosomal proteins, tRNA genes and miRNA families on *C. avellana* cv Tombul chromosomes. **B.** Distribution of predicted mature miRNA lengths. **C.** Copy numbers of miRNA families, in the genome and in tandem repeat clusters.

in the table, but experimental confirmation, e.g. from small RNA sequencing, should be carried out to confirm which are produced *in vivo*. Predicted mature miRNAs ranged from 18-24 nt in length with the majority being 21 nt (Fig. S4B). The 153 candidate miRNA genes included members of 52 miRNA families, of which 26 had multiple copies in the genome (Supplementary Fig. 4C). Tandem duplications contributed to the expansion of these miRNA families to a variable degree; for example, most miR171 genes were found as single loci distributed between 5 different chromosomes, whereas all 12 miR399 genes were found in a single cluster on pchr02. These structural variations may provide clues as to the contrasting development

of these miRNA families during the evolution of *Corylus*, and also to their current regulation and functions.

Potential miRNA targets were predicted based on the presence of their complementary target sequences in Tombul transcriptome sequences, which were then annotated with GO terms (Table S11).

2.8 Protein-coding genes and annotation

Prediction of protein-coding genes and annotation with MapMan bins, GO terms and conserved protein domains was carried out as described in the main text, and results are summarized in Supplementary Tables 12+. With respect to the 3 main GO categories, Biological Process GO terms were assigned to 15,425, Cellular Component terms to 15,313, and Molecular Function terms to 16,473 gene models. For each gene model, the most specific (highest level) GO terms assigned for each category are recorded in Supplementary Table 12. The most abundant conserved InterPro protein signatures were derived from a variety of protein kinases; e.g. 1401 gene models were annotated with IPR000719, 'protein kinase catalytic domain', accounting for almost 5% of all high-confidence gene models; this is similar to the abundance observed in other sequenced plant genomes [51] and indicates the importance of signal transduction in controlling many aspects of plant biology. Apart from highly prevalent domain structures that are found in proteins with diverse functions (e.g. NAD(P)-binding, Pentatricopeptide repeat, Zinc finger, WD40, and Armadillo domains) other notable abundant domains included the NB-ARC motif, found in 279 gene models, the large majority of which also contained leucine-rich repeats (LRRs). These genes belong to the NB-LRR family of pattern recognition receptors, which play a critical role in the plant response to pathogens. Homeobox-like and Cytochrome P450 domains were each found in over 300 gene models, making these the most abundant transcription factor and enzyme families respectively in the hazelnut genome.

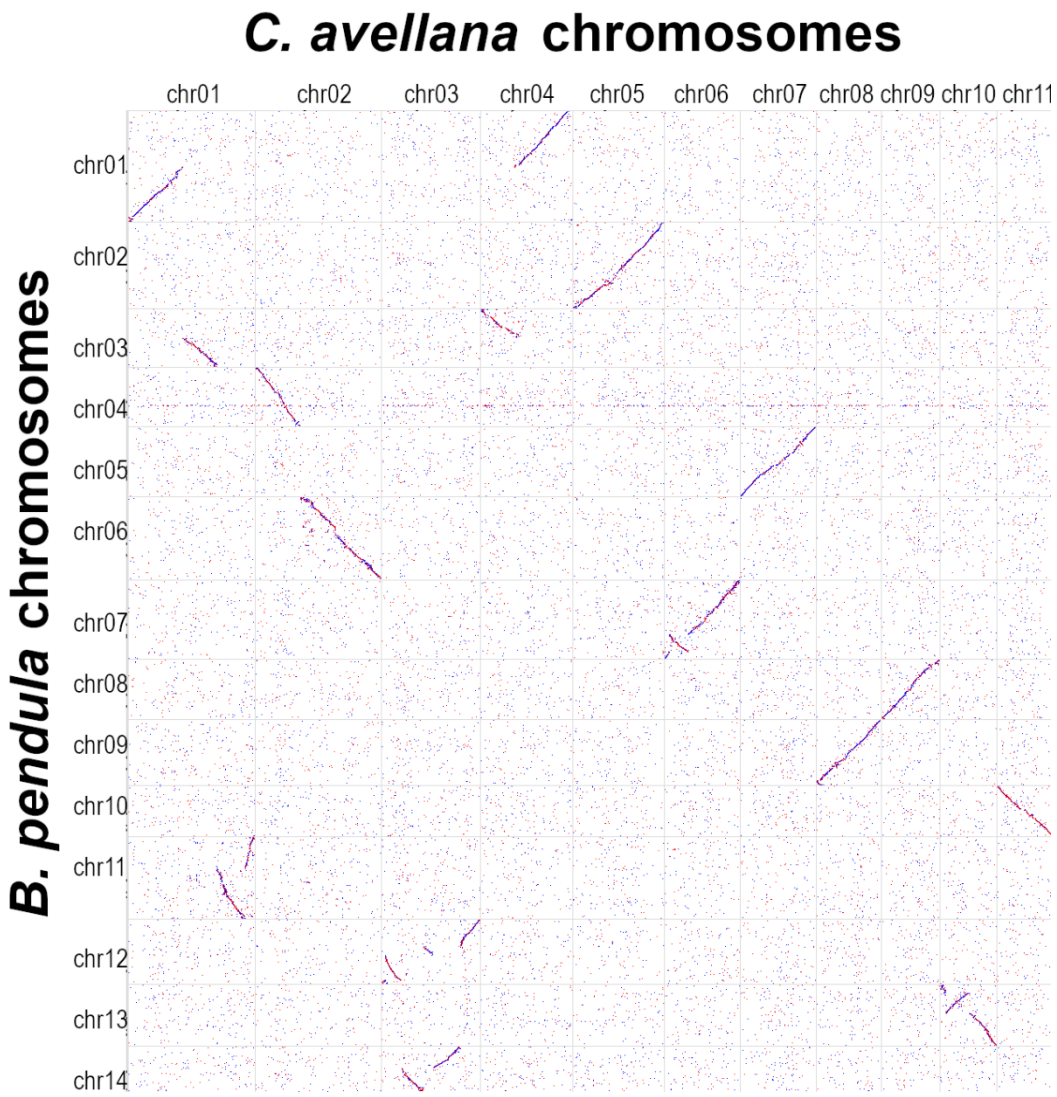


Figure S5. Dot plot representing whole genome alignment of the *C. avellana* and *B. pendula* chromosome assemblies. Blue dots indicate sequence matches in the same orientation, red dots indicate inverted sequence matches.

2.9 Comparative Genome Alignment

The chromosome pseudomolecule sequences from *C. avellana* var. Tombul were aligned with those of *B. pendula*[32] using LAST, with masking of short repeats and other parameters as recommended by the software authors (https://github.com/mcfrith/last-genome-alignments). A dot-plot of the alignment is shown in Figure S5. Strikingly, *C. avellana* pchr05, 07, 08, 09 & 11 are collinear with *B.pendula* chromosomes 2, 5, 9, 8 & 10 respectively along their entire length. pchr06 and pchr10 are largely collinear with *B. pendula* chromosomes 7 & 13 respectively,

with a single large inversion in each. The greatest structural difference between the species is seen in *C.avellana* pchr03, which includes 6 large segments that are rearranged into *B. pendula* chr12 & chr14. pchr02 is split into *B. pendula* chr4 & chr6, while pchr01 & 04 are split into 6 large segments in total, pairs of which are assorted into *B. pendula* chromosomes 1, 3 & 11. The observed structural differences can be explained parsimoniously by assuming an ancestral Betualceae karyotype of 17 chromosomes, with 6 and 3 chromosome fusion events taking place in the *Corylus* and *Betula* lineages respectively, along with 6 intra-chromosomal inversions distributed between the lineages.

2.10 Characterization of the *MLO* gene family

MLO proteins are located at the plasma membrane and have a well-conserved structure consisting of 7 transmembrane (TM) helices, followed by a C-terminal cytoplasmic Calmodulin binding domain, making them relatively easy to identify by sequence homology. Based on the initial automated gene annotation described above, 24 gene models were annotated with the MLO-related protein domain (IPR004326). However, over half of these gene models had a predicted protein product that was significantly shorter or longer than known MLO proteins (<350 or >650 amino acids). Therefore all of the gene models were manually re-annotated, using blastp against non-redundant plant proteins confirm their identity, tblastn to identify regions of MLO homology that had been omitted from gene models, and FGENESH[52] to re-evaluate the gene models predicted by Augustus[25]. Where FGENESH and Augustus gave conflicting predictions, the helical structure of both predicted proteins was assessed using TMHMM[53], and the one most consistent with known MLO protein topology selected. The results of re-annotation are summarized in Table S16. After re-annotation, 11 gene models were retained as encoding full-length MLO proteins; the remaining genes mostly seem to be truncated or disrupted duplicates of the full-length *MLO*s, and so were labelled as '*MLO*-like genes'.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Table S16: *MLO*-related gene model predictions and results of re-annotation.

Gene Model ID	Length (aa)	Best blastp match in *Arabidopsis*	Result of re-annotation
Cav01g18310	140	AtMLO4	Merged with upstream, un-annotated ORF to give full-length *MLO* gene
Cav02g00450	573	AtMLO10	Verified as full-length *MLO* gene
Cav02g01800	306	AtMLO4	Two separate, truncated genes, probably a degraded ancestral *MLO* gene
Cav02g01810	202		
Cav02g28680	484	AtMLO12	Original gene model had only 422 amino acids; replaced with better FGENESH gene model
Cav02g28690	158	AtMLO12	Truncated ortholog of Cav02g28680
Cav02g28700	168	AtMLO6	Truncated ortholog of Cav02g28680
Cav02g28710	180	AtMLO12	Truncated ortholog of Cav02g28680
Cav03g23960	586	AtMLO11	Original gene model had only 537 amino acids; replaced with better FGENESH gene model
Cav04g11090	552	AtMLO1	Verified as full-length *MLO* gene
Cav04g11100	272	AtMLO1	Fragmented ortholog of Cav04g11090, disrupted by an LTR/retroviral insertion
Cav04g11110	247		
Cav07g08600	518	AtMLO1	Verified as full-length *MLO* gene
Cav07g19540	590	AtMLO6	Verified as full-length *MLO* gene
Cav07g19550	348	AtMLO6	Truncated ortholog of Cav07g19540
Cav07g19560	603	AtMLO6	Full-length *MLO,* near ortholog of Cav07g19540
Cav07g19570	106	AtMLO6	Truncated ortholog of Cav07g19540
Cav07g19580	798	AtMLO2	Possibly 2 ancestral *MLO* genes spliced together
Cav07g19590	541	AtMLO12	Original gene model had only 474 amino acids; replaced with better FGENESH gene model
Cav08g17470	403	AtMLO12 (part)	Truncated ortholog, similar to fragment upstream of Cav02g00450
Cav08g17510	520	AtMLO5	7-8 TM helices, closest to Cav02g00450
Cav09g12130	532	AtMLO12	Truncated ortholog (of Cav10g09700?)
Cav09g12140	612	AtMLO3	Original gene model had only 345 amino acids; replaced with better FGENESH gene model
Cav10g09690	427	AtMLO12	*MLO*-like, N-terminal truncated
Cav10g09700	589	AtMLO2	FGENESH gene model, very slightly better than original prediction (587 amino acids)
Cav10g09710	192	AtMLO3	Two separate, truncated genes, probably a degraded ancestral *MLO* gene
Cav10g09720	318		

Multiple sequence alignment of the *C. avellana* MLO genes with those from *A. thaliana* & *Malus domestica* (Figure S6) shows that the TM helix sequences are usually highly conserved, with some insertions in individual genes in the loops between them, and greater diversity at the C-terminus. However, Cav07g08600 seems highly divergent from the other genes in part of the TM region (between 600-750 in Fig. S6); this gene was also only predicted to contain 5 TM

helices instead of 7. Therefore further structural or functional studies are required to confirm whether it encodes an intact MLO protein.

In addition to the alignment of full-length genes mentioned above, full-length *MLO* and *MLO*-like genes from Clades V, VI & VII were aligned with orthologs from the other tree genomes used for evolutionary analysis. Phylogeny within these groups was reconstructed using the Maximum Likelihood method, allowing variable substitution rates between sites (Figure S7). In cases in which an historic *MLO* gene has been degraded into 2 smaller putative genes, the amino acid sequences of the 2 fragments were merged with a 'XXXXX' linker in between, to provide a more complete alignment (e.g. Cav02g28690 & Cav02g28700, Fig. S7A). Based on this reconstruction, while most other *MLO* genes are evenly represented in the tree genomes examined here, both Cav02g28680 (Fig. S7A) and Cav07g19540 (Fig. S7B) appear to have undergone repeated gene duplications that are unique to the *Corylus* lineage.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

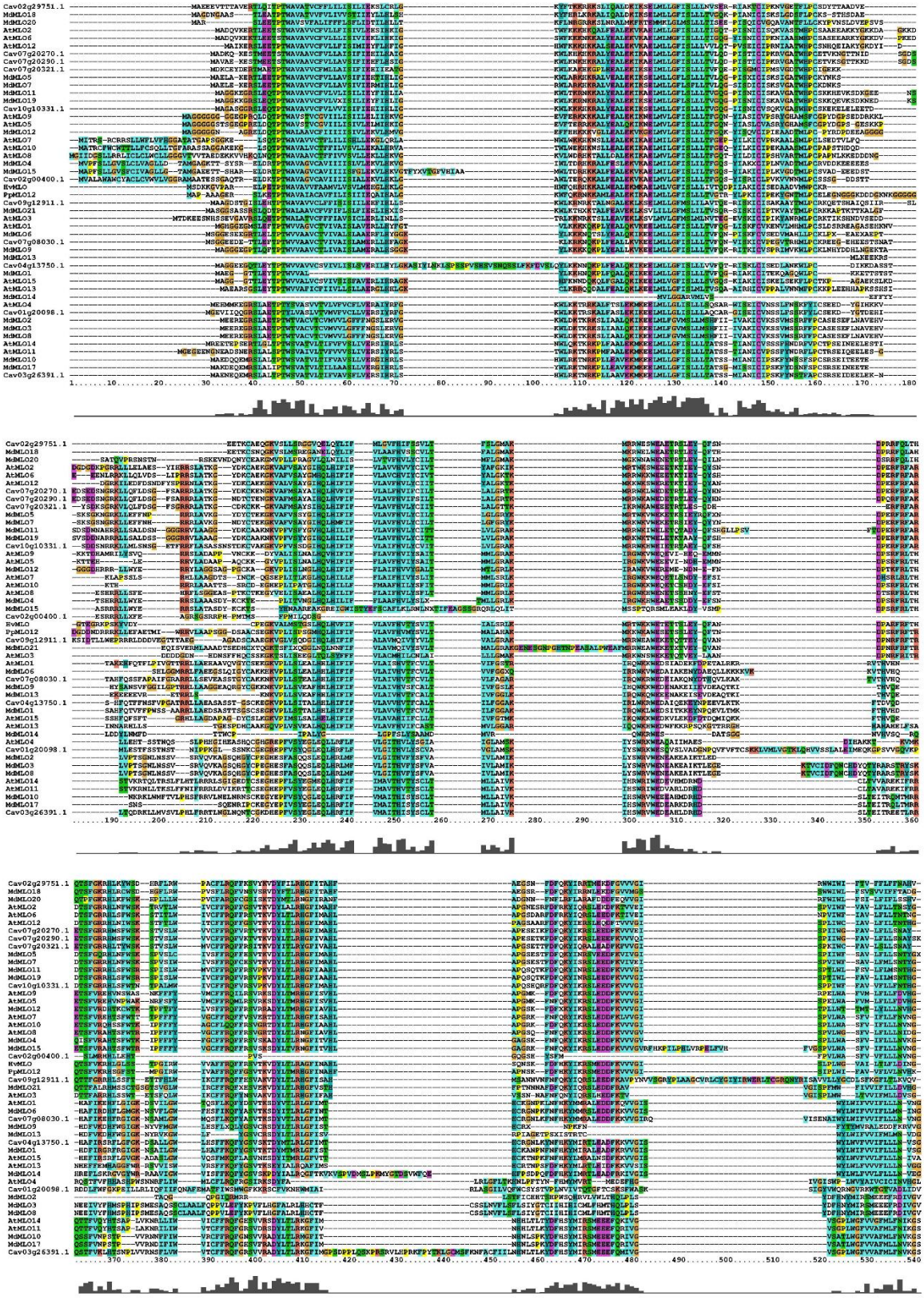


Figure S6. Multiple sequence alignment (1-540 aa) of predicted MLO protein sequences from *C. avellana*, *A. thaliana* and *M. domestica*, produced using MUSCLE with UPGMA clustering.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

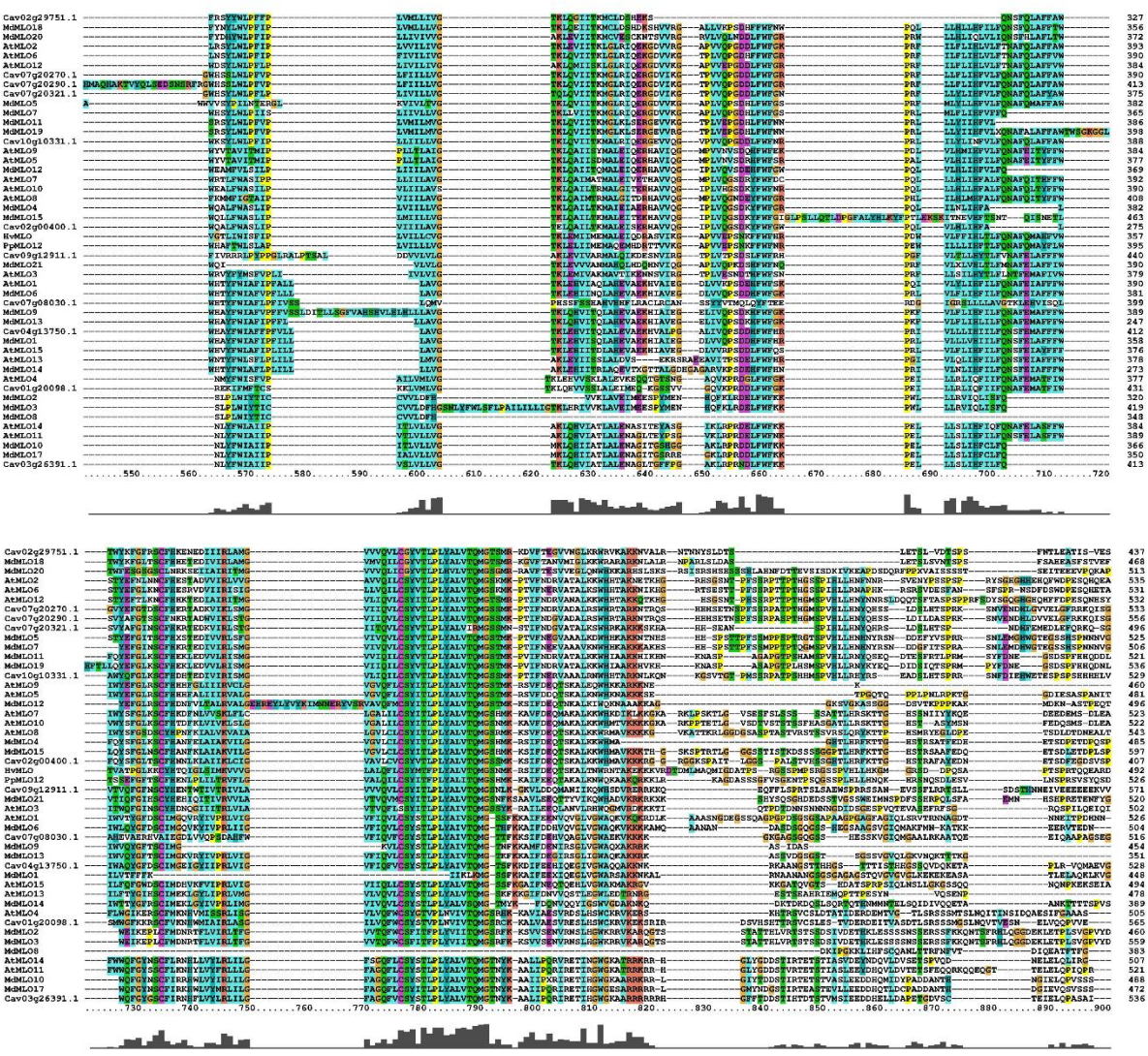


Fig. S6 (continued). Multiple sequence alignment (541-1080 aa) of predicted MLO protein sequences from *C. avellana, A. thaliana* & *M. domestica*. Continued from previous page.

Supporting Information

"A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement" Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

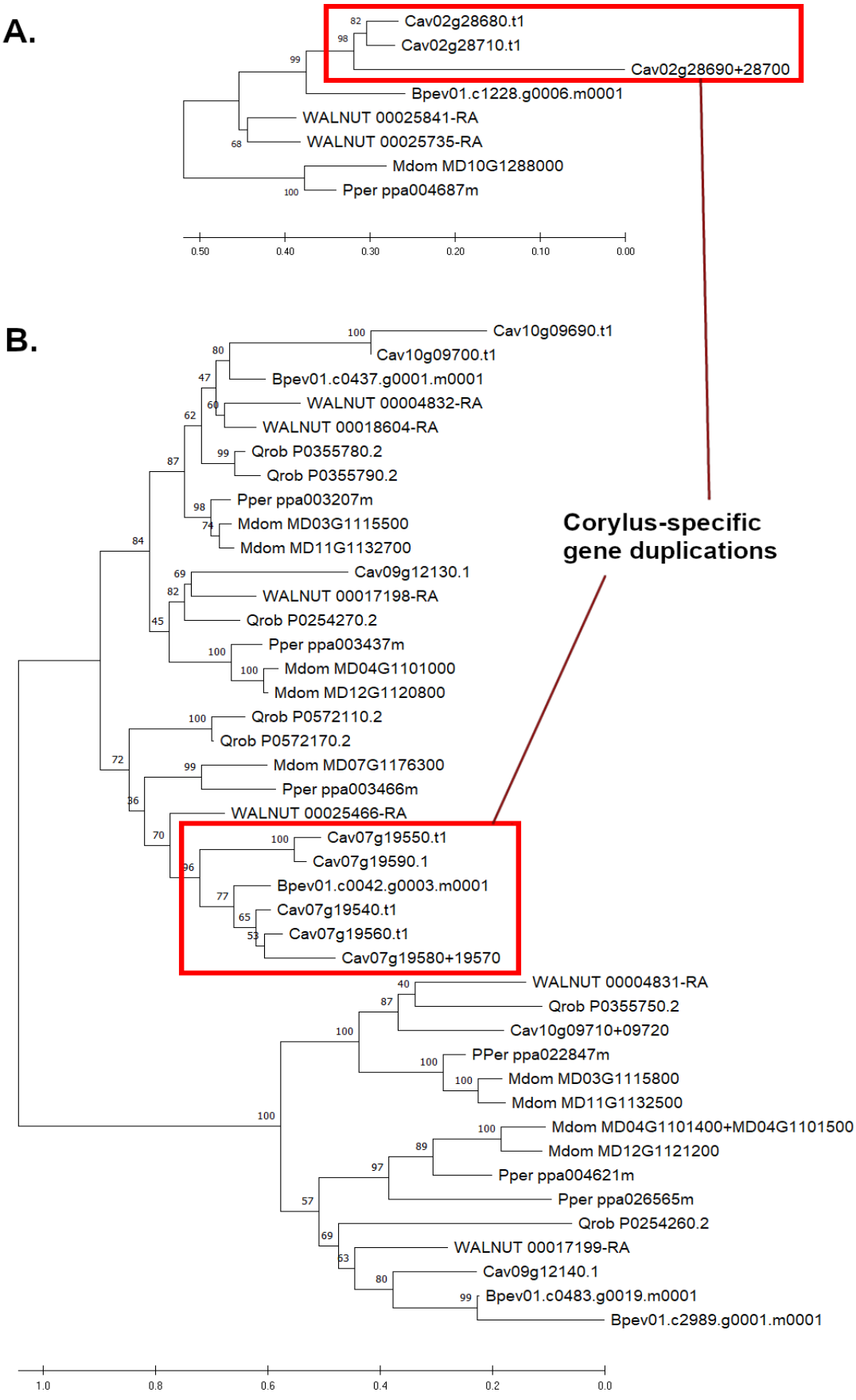


Figure S7. Phylogenetic inference (ML) of tree *MLO*-like genes. **A.** Clade VII, **B.** Clade V + VI

2.11 Identification of Allergen-coding genes and homologs

Gene models encoding most of the reported hazelnut allergens and thier homologs were easily identified from the predicted proteins by sequence similarity. The seed storage proteins (Cor a 9, Cor a 11, Cor a 14) had been identified during *de novo* gene modelling but were not represented in our high-confidence protein set, owing to their absence from our transcriptome data; this is attributed to their exclusive expression in seeds. Complete protein sequences were not available for Cor a 15 or Cor a TLP. For Cor a 15, the genome only contained 4 putative oleosin genes, 2 of which (Cav05g25120 & Cav08g12860) were unambiguously identified as *Cor a 12* and *Cor a 13* respectively. Cav03g05140 encodes a predicted oleosin with a conserved central hydrophobic domain but different N- and C- terminal sequences from Cor a 12, consistent with the description of Cor a 15 [54], whereas Cav06g12950 encodes a putative novel homolog of Cor a 13.

In the case of Cor a TLP, the only sequence information available was that of 2 peptides identified by Edman sequencing[55]. These peptides were both used to search the predicted hazelnut proteins. In total 48 putative TLPs were predicted in the *C. avellana* genome, with a large cluster of 19 highly conserved gene models located in the sub-telomeric region of the short arm of pchr09. One of the peptides matched all of these predicted TLP proteins near their conserved N-terminus with 88% identity; however, the second peptide only gave a partial match (10/17 amino acids) near the C-terminus of some members of the same cluster. Of these the best match was to Cav09g00690, which also encoded the TLP with the highest sequence identity (76.42%) to the known peach TLP allergen Pru p 2.01. Six other genes from the same cluster also matched both Pru p 2.01 and Mal d 2 with >75% identity, and are listed as probable cross-reactive allergens in Table S17.

Supporting Information

“A chromosome-scale genome assembly of European Hazel (Corylus avellana L.) reveals targets for crop improvement” Lucas SJ, Kahraman K, Avşar B, Buggs RJA, Bilge I (2020)

Supplement References