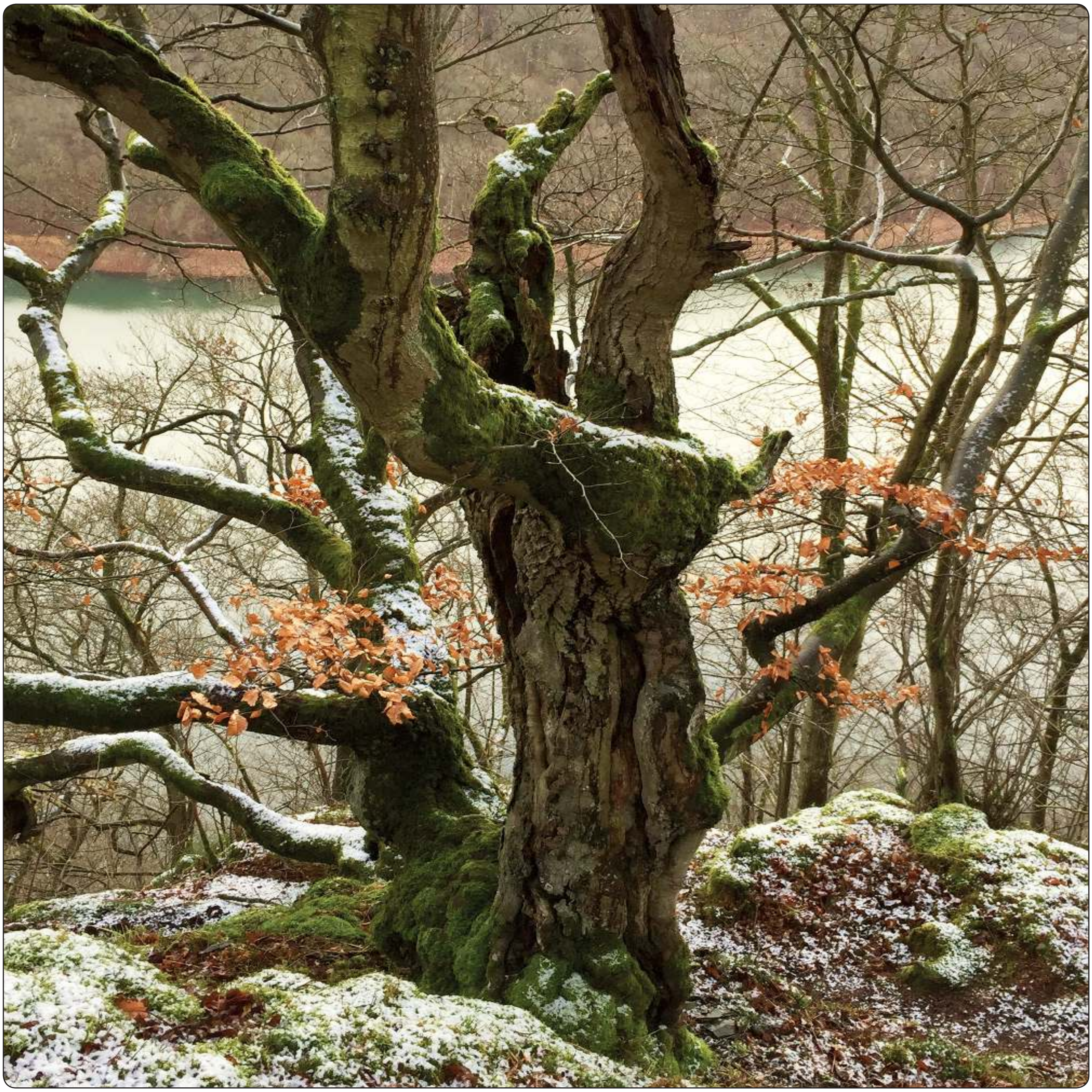
(GIGA)n SCIENCE



The sequenced individual Bhaga at the time of sampling. Note the very low branching on the cliff, with a major part of the individual reaching over the edge

Mishra et al.

OXFORD
UNIVERSITY PRESS

Mishra et al. *GigaScience* 2018, **7**:6
https://doi.org/10.1093/gigascience/giy063

GigaScience, 7, 2018, 1–8

doi: 10.1093/gigascience/giy063
Advance Access Publication Date: 28 May 2018
Data Note

DATA NOTE

A reference genome of the European beech (*Fagus sylvatica* L.)

Bagdevi Mishra[1,2], Deepak K. Gupta[1,2], Markus Pfenninger[1,3], Thomas Hickler[1,4], Ewald Langer[5], Bora Nam[1,2], Juraj Paule[6], Rahul Sharma[1], Bartosz Ulaszewski[7], Joanna Warmbier[7], Jaroslaw Burczyk[7] and Marco Thines[1,2,*]

[1]Senckenberg Biodiversity and Climate Research Centre (BiK-F), Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany, [2]Goethe University, Department for Biological Sciences, Institute of Ecology, Evolution and Diversity, Max-von-Laue-Str. 9, D-60438 Frankfurt am Main, Germany, [3]Johannes Gutenberg Universität, Fachbereich Biologie, Institut für Organismische und Molekulare Evolutionsbiologie (iOME), Gresemundweg 2, 55128 Mainz, [4]Goethe University, Department for Geology, Institute of Geography, Max-von-Laue-Str. 23, D-60438 Frankfurt am Main, Germany, [5]University of Kassel, FB 10, Department of Ecology, Heinrich-Plett-Str. 40, D-34132 Kassel, Germany, [6]Senckenberg Research Institute and Natural History Museum Frankfurt, Department of Botany and Molecular Evolution, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany and [7]Kazimierz Wielki University, Department of Genetics, ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland

[*]**Correspondence address.** Marco Thines, oethe University, Department for Biological Sciences, Institute of Ecology, Evolution and Diversity, Max-von-Laue-Str. 9, D-60438 Frankfurt am Main, Germany, E-mail: m.thines@thines-lab.eu http://orcid.org/0000-0001-7740-6875

Abstract

Background: The European beech is arguably the most important climax broad-leaved tree species in Central Europe, widely planted for its valuable wood. Here, we report the 542 Mb draft genome sequence of an up to 300-year-old individual (Bhaga) from an undisturbed stand in the Kellerwald-Edersee National Park in central Germany. **Findings:** Using a hybrid assembly approach, Illumina reads with short- and long-insert libraries, coupled with long Pacific Biosciences reads, we obtained an assembled genome size of 542 Mb, in line with flow cytometric genome size estimation. The largest scaffold was of 1.15 Mb, the N50 length was 145 kb, and the L50 count was 983. The assembly contained 0.12% of Ns. A Benchmarking with Universal Single-Copy Orthologs (BUSCO) analysis retrieved 94% complete BUSCO genes, well in the range of other high-quality draft genomes of trees. A total of 62,012 protein-coding genes were predicted, assisted by transcriptome sequencing. In addition, we are reporting an efficient method for extracting high-molecular-weight DNA from dormant buds, by which contamination by environmental bacteria and fungi was kept at a minimum. **Conclusions:** The assembled genome will be a valuable resource and reference for future population genomics studies on the evolution and past climate change adaptation of beech and will be helpful for identifying genes, e.g., involved in drought tolerance, in order to select and breed individuals to adapt forestry to climate change in Europe. A continuously updated genome

Received: 23 January 2018; **Revised:** 20 April 2018; **Accepted:** 19 May 2018

browser and download page can be accessed from beechgenome.net, which will include future genome versions of the reference individual Bhaga, as new sequencing approaches develop.

Keywords: forest tree; *fungi*; *genomics*; *hardwood*; *hybrid assembly*; *transcriptomics*

Data Description

Context

European beech (*Fagus sylvatica* L., NCBI Taxon ID: 28 930) is one of the most important and widespread broad-leaved tree species in Europe. Its natural range extends from southern Italy to southern Scandinavia and from the Iberian Peninsula to Crimea [1]. Under favourable conditions, in particular in Central Europe, it can outcompete all other tree species and form monospecific stands in which, due to shading, other broad-leaved species can hardly establish [2]. Because of their cultural and environmental importance, as well as their global uniqueness, ancient and primeval beech forests in Europe, with five areas located in Germany, have been listed as UNESCO World Heritage sites [3]. Langer et al. [4] analyzed the species composition of these forests and concluded a need for conservation of near natural or primeval beech forest stages for their richness in fungal species.

There have been 1,766 fungal species reported associated with beech, ranging from general commensals to specialised pathogens and symbionts, such as the very common obligate mycorrhizal symbiont *Lactarius blennius* (beech milkcap), with a distribution corresponding to the natural distribution of beech [5, 6]. On average, 25 fungal species are associated with the dead wood of *F. sylvatica* [7]. Among them are threatened species and species with natural value such as *Hericium coralloides* and *Phleogena faginea* [8, 9]. Nitrogen uptake by beech roots is also highly dependent on the mycorrhizal community [10]. Thus, the European beech is in intimate contact with a variety of fungi.

Even though its natural area of dominance [11] has been reduced by land use and planting other commercially important species, such as Norway spruce (*Picea abies*; [12]), European beech remains an important hardwood species on the European scale. However, as European beech does not cope very well with dry and hot conditions, fire, and flooding, its suitability under a potentially more extreme climate in the future is debatable [13]. Thus, genetic and genomic data are crucial for understanding its adaptive capacity, in particular, under climate change [14], which will also lead to a change in biotic stress, including fungal pathogens [15, 16].

Several tree genomes have been released over the past decade, among them oaks [17, 18] and Chinese chestnut [19] of the beech family (*Fagaceae*). However, despite its economic and ecological importance, genetic and genomic resources in the genus *Fagus* (beeches) are limited to some studies on the genetic diversity and candidate genes using single-nucleotide polymorphism data [20-23], a few genome-wide associations studies [24, 25], investigation into methylation patterns [26], and some transcriptome data [27, 28]. Thus, it was our aim in this study to provide a draft assembly of the European beech and to make it available to the research community for in-depth analyses and follow-up studies, taking advantage of the genomic resource. The risk of contamination with a variety of microorganisms, including bacteria and the numerous fungi found in association with trees in general and beech in particular [29], is high when conducting sampling of specimens from nature, as evidenced by the large amount of contaminant DNA in the effort of sequencing the olive tree genome from a 1,000-year-old individual [30]. Thus, we are also describing a method of DNA extraction from dormant buds that, in our case, led to the absence of contaminant organisms in the assembly.

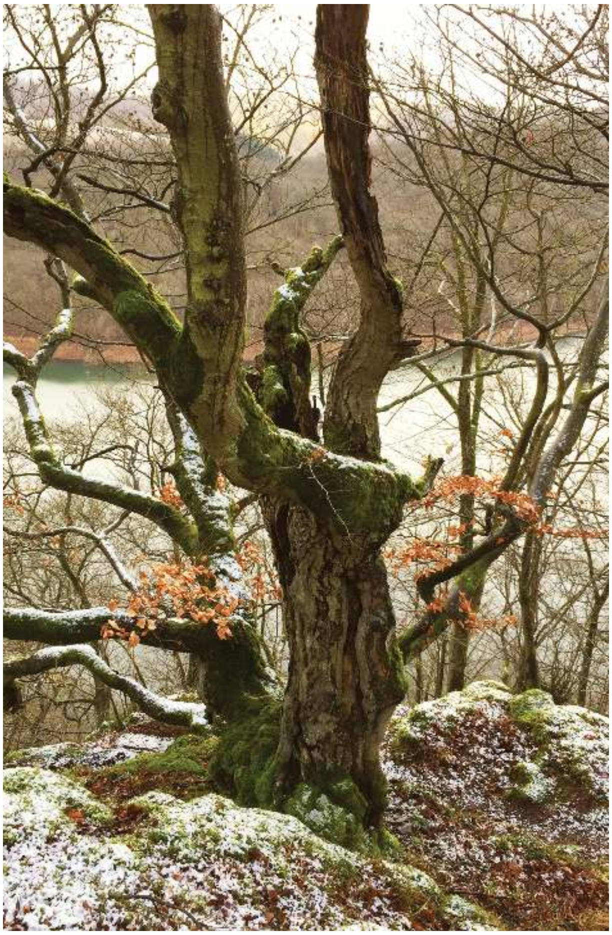


Figure 1: The sequenced individual Bhaga at the time of sampling. Note the very low branching on the cliff, with a major part of the individual reaching over the edge.

Methods

Selection of the sequenced individual

For the genome sequencing, an individual tree standing on a rocky outcrop on the rim of a scarp to the Edersee (German Kellerwald-Edersee National Park) was selected (Fig. 1). The individual, named Bhaga (the reconstructed common root of the common name of the tree in several European languages), is estimated to be up to 300 years old, based on its poor stand, low branching, as well as bark and stem characteristics. A direct measurement was not possible because the trunk is not fully preserved due to the tree's age. An old individual was selected to avoid the influence of modern forestry on the tree's genetic makeup .

Flow cytometric genome size and nucleotide composition

Relative genome size and absolute genome size was estimated by flow cytometric analyses of fresh leaf buds using a CyFlow space (Partec, Münster, Germany). Leaf buds (without bud scales) of the analysed sample and leafs of the internal standard (Glycine max cv. "Polanka" (2C = 2.50 pg) were treated and analysed as described previously [31].

DNA and RNA extraction

A modified protocol based on the standard CTAB (cetyl trimethylammonium bromide) method described by Doyle and Doyle [32] was applied. The CTAB extraction buffer consisted of 100 mM Tris-HCl, 20 mM Ethylenediaminetetraacetic acid (EDTA), 1.4 M NaCl, 2% CTAB, 0.2% ß-mercaptoethanol, and 2.5% polyvinylpyrrolidone. For DNA extractions, about 100 buds (collected in February 2015) with a few millimeters of the subtending branchlets were cut from twigs of a larger branch, transported to the laboratory on ice, and surface sterilized by gentle shaking for 2 minutes in 4% sodium hypochlorite solution containing 0.1% Tween. Subsequently, the buds were rinsed with sterile water until no foam formation was evident. Then, the water was poured off and the buds were descaled after cutting off the subtending branchlet with sterile scalpels. The dormant leaf tissue in the buds was ground in liquid nitrogen using a mortar and pestle. A total of 1,200 mg of powdered tissue was distributed to 24 2-mL reaction tubes. Each sample was thoroughly mixed with three 3-mm metal beads in 600 μL of extraction buffer and incubated at 60°C for 30 minutes. After this, 600 μL of phenol:chloroform:isoamyl alcohol (25:24:1) (PCI) was added, and the tubes were gently mixed by inversion. Subsequently, the tubes were centrifuged at 19,000 *g* for 2 minutes. Next, 500 μL of the supernatant were transferred to a new tube and 600 μL of PCI was added. The tubes were centrifuged again for 2 minutes, and each 500 μL of the supernatant transferred to a new tube. Subsequently, 15 μL RNase A solution (100 mg/mL) were added to each tube, and the tubes were incubated at 37°C for 30 minutes. After the incubation, 600 μL of chloroform was added, and the tubes were gently shaken. Subsequently, the tubes were centrifuged at 19,000 *g* for 2 minutes. The supernatant of all tubes was transferred to a 45-mL reaction tube. Then, 3 M sodium acetate solution at pH 5.3 (supernatant to 3 M sodium acetate solution = 1:0.09) and 100% ethanol (supernatant to ethanol = 1:2) were added to the supernatant, and the tube was gently mixed by inversion. Afterward, it was incubated at –20°C for 30 minutes and centrifuged at 4800 *g* for 3 minutes at 4°C. The supernatant was carefully poured off, and the pellet was washed twice with 70% ethanol. After a final centrifugation at 4800 *g* for 2 minutes at 4°C, the supernatant was poured off carefully, and the pellet was dried at room temperature in a clean laminar flow bench for approximately 1 hour. Subsequently, the pellet was dissolved in prewarmed (40°C) 0.1 x Tris-EDTA buffer for further analysis. RNA was isolated from ground dormant leaf tissue and prepared as described above using a NucleoSpin RNA Plant Kit (Macherey-Nagel, Düren, Germany) according to the protocol supplied with the kit. The extracted DNA and RNA were checked for integrity and quantity using agarose gel electrophoresis and fluorometry on a Qubit v3 device (ThermoFisher, United States), respectively.

Sequencing

From genomic DNA shotgun TruSeq paired-end libraries of 300 bp and 600 bp insert lengths and long-jumping-distance (LJD) libraries of 3 kbp, 8 kb, and 20 kb were constructed for paired-end sequencing (2 × 100 bp) on an Illumina HiSeq 2000 Sequencer Iillumina, United States) by a commercial sequencing provider (LGC Genomics GmbH, Germany). In addition, libraries with a target insert size of 20 kb for Single Molecule Real-Time (SMRT) sequencing on a Pacific Biosciences (PacBio) RSII instrument (United States) using the DNA/Polymerase Binding Kit P6 were constructed and sequenced by a commercial sequencing provider (Eurofins Genomics, Germany) using 6 SMRT cells. In addition, both mRNA-enriched and ribosome-depleted TruSeq paired-end libraries were prepared and subsequently sequenced on a HiSeq 2000 instrument by LGC Genomics GmbH (Germany).

Assembly and quality control

Illumina reads were checked for adapter sequences and bad-quality read ends using Trimmomatic v0.36 (Trimmomatic, RRID:SCR_011848) [33] with the following parameters: "TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:70". Reads with Ns in the sequences were filtered using Sickle (version 1.33) [34]. The final cleaned dataset used included reads with an average quality of more than 30, were longer than 70 bp, and were without Ns. The PacBio reads were corrected by the filtered Illumina reads using Proovread (version 2.14.0) [35], and the corrected reads were further used for the assembly.

All sequencing data as well as the genome assembly can be found under accession number PRJEB24056 at the European Nucleotide Archive (ENA) [36]. The assembly was done using a hybrid assembly approach in which an initial assembly was built using Velvet v.1.2.10 [37] on shotgun reads with insert lengths of 300 bp and 600 bp (35 Gb, corresponding to 75x coverage after adapter trimming and filtering) with a k-mer length of 63 and without scaffolding. This pre-assembly of 360 Mb with a minimum contig length of 300 bp was taken as a base for a DBG2OLC (last update, June 11, 2015) [38] hybrid assembly using corrected PacBio reads >150 nucleotides (7.9 Gb, corresponding to 17x coverage, mean size 9,487 nucleotides, median 9,162 nucleotides, longest sequence 47,053 nucleotides) with a k-mer length of 17, a k-mer matching threshold for each contig of 5, minimum matching k-mers for each two reads of 30, adaptive k-mer threshold for each contig of 0.002, and chimera removal option set to 1. The resulting assembly of 541 Mb was further scaffolded with Illumina LJD libraries using SSpace (basic version) (SSPACE, RRID:SCR_005056) [39]. The genome size was estimated using k-mer counting based on the depth distribution as computed by Jellyfish v 2.0 (Jellyfish, RRID:SCR_005491) [40] using 15-mers and considering all coverage depths using R-scripts.

A CEGMA v 2.5 (CEGMA, RRID:SCR_015055) [41] analysis was performed to test for the completeness and continuity of the beech genome assembly, along with other published tree genomes. In addition, the assembly was evaluated with plant-specific Benchmarking Universal Single-Copy Orthologs (BUSCO, RRID:SCR_015008) [42].

Gene prediction

Splice alignments of Illumina RNA sequencing (RNA-seq) data (filtered using the same criteria as above for genomic reads, in total 3.2 Gb) were built using Tophat2 v 2.0.10 (TopHat, RRID:SCR_013035) [43] using the draft genome. This alignment was used in Blast2GO v4.1 (Blast2GO, RRID:SCR_005828) [44] along with a pretrained dataset from *Arabidopsis thaliana*. Genes were predicted on both strands. Genes with a length of more than 90 nucleotides with both a start and a stop codon were considered. For the other parameters, default values were opted. Genes were annotated using Blast2GO. For the sequence similarity-based annotation, a locally downloaded protein-RefSeq database [45] was queried using the Blastp-fast algorithm of the Basic Local Alignment Search Tool (BLAST), version 2.2.30+ (National Center for Biotechnology Information [NCBI] BLAST, RRID:SCR_004870). In a second, less-stringent approach to predict more splice variants, splice-alignment information from RNA-Seq mapping was used along with the single-copy protein sequences predicted in the BUSCO pipeline [42], in the BRAKER2 pipeline (version 2.1.0) [46] using GeneMark-ET v 4.29 [47] and Augustus v3.2.6 (Augustus: Gene Prediction, RRID:SCR_008417) [48]. The splice alignments of RNA-seq reads mapped on the genome were also used as extrinsic evidence in this approach.

Repeat prediction

RepeatScout v1.0.5 (RepeatScout, RRID:SCR_014653) [49] was used for *de novo* identification of repeat elements and for generating a repeat element database. This database was used in RepeatMasker v4.0.5 (RepeatMasker, RRID:SCR_012954) [50] to predict repeat elements. Putative repeats were further filtered on the basis of their copy numbers. Those repeats represented with at least 10 copies in the genome were retained.

General genomic features

For each annotated gene, the shortest distance to the next gene on the same scaffold was measured. In addition, the distance between all heterozygous sites was assessed, as identified by positions with a two-base ambiguity code in the assembly. For this, genomic reads were mapped using MAQ (version 3) [51], and positions were scored as heterozygous if the frequency of the lesser base was at least 40%. For the aforementioned analyses, the assembly was divided into nonoverlapping windows of 10-kb size. For each of the resulting 50,994 windows, gene density, GC-content, and genetic diversity were determined. Exon density was measured as the proportion of each window annotated as protein-coding and GC content as proportion of G and C bases. Genetic diversity was approximated by the proportion of heterozygous sites in each window. The values were extracted from the assembly and GFF files using custom-made Python scripts, available upon request. Because genome windows in spatial proximity may not represent independent data, each parameter was tested for spatial autocorrelation using Moran's I as test statistics. The relations between the parameters were explored using linear regression models.

Screening for contamination

The genic regions of *F. sylvatica* were blasted against two databases, one containing genes from *Arabidopsis thaliana* and the other containing genes from *Fungi* and *Straminipila*, using an e-value cutoff of $10e^{-5}$ and extracting the top hits. The genic regions having a fungus as the top hit were blasted against the non-redundant database from NCBI [52] to reveal whether these were indeed specific to fungi. Local alignments of the genic regions remaining after this filtering process to the supposed fungal homologs were subsequently manually inspected for the distribution of conserved features.

In addition, the assembled genome was chopped into 300-bp fragments and subjected to analysis with MEGAN (version 5) [53]. The fragmented genome was blasted against the nucleotide database downloaded from NCBI using an e-value cutoff $10e^{-8}$ and a 70% identity cutoff.

Data description, validation, and control

Genome summary

Raw reads, assembly, and annotations are available from the ENA at accession number PRJEB24056 and at the Beech Genome Resource website [54]. The genome size was estimated to be 541 Mb based on 15-mer counts (Supplementary Fig. S1), while the draft genome assembly was 542 Mb. The assembly comprised 6,491 scaffolds, with 0.12% Ns. The largest scaffold was 1.15 Mb, the N50 length was 145 kb, and the L50 count was 983. Also, 58.36% of the genome is classified as interspersed repeats and around 2% of the genome is comprised of simple repeats. The locations of the interspersed repeats and the simple repeats in the scaffolds are provided as a gff file for download and as a separate track in the genome browser [54]. In total, 62,012 genes and 73 splice variants were predicted using Blast2Go, of which 58,211 genes had received at least one RNA-seq read support (50,723 genes were supported by at least five reads). The average number of exons per gene was 4.59, and the distribution of the number of exons per gene was similar to that of other genomes (Supplementary Fig. S2). The BRAKER2-based gene prediction resulted in 100,822 complete genes, including 1,332 splice variants. Of the genes predicted by BRAKER2, 90,936 genes were supported by at least one RNA-seq read (73,598 genes were supported by at least five reads). This gene set is given as an additional track in the genome browser and as a supplementary gene annotation file on the genome resource page [54]. A total of 60,879 genes predicted by Blast2GO gene were found to be present in the gene set predicted by BRAKER2 pipeline according to a homology-based sequence similarity analysis using Blastp (version: 2.2.29+) with an e-value cutoff of 10e-10.

The mean (median) minimum observed distance between annotated genes on the same scaffold was 2,696 (1,617) bp, ranging from 1 bp to about 73 kb (Supplementary Fig. S3). The mean (median) distance among neighboring heterozygous sites was 460 (95) bp, with a range of 1 to 136 kb (Supplementary Fig. S4). Gene density in 10-kb windows was between 0 and 0.99 coverage, with a mean (median) of 0.196 (0.170) (Fig. 2A). The respective density values for exons fell to between 0 and 0.87, with a mean of 0.196 and a median of 0.170. The mean (median) GC content of the windows was 0.356 (0.349; Fig. 2B). This is about 5% lower than published values for beech [55] but refers here to only the nonrepetitive regions of the genome. On average, 2 in 1,000 sites were heterozygous (0.0019), with a range of 0 to 0.021.

Because there was no spatial autocorrelation among adjacent nonoverlapping 10-kb windows or multiples of it (Moran's I <10–4) for either parameter, we could treat the extracted values as independent data points. There was a very strong relationship between exon density and GC content ($r^2 = 0.91$, $P < 0.0001$; Fig. 3A), while the correlation between gene density and GC content was marginal ($r^2 = 0.02$, $P < 0.0001$). This pattern was observed in many angiosperms and is usually explained as GC-biased gene conversion [56].

Positive, purifying, and background selection on functional genome elements is thought to negatively influence genetic diversity [57]. Therefore, a negative correlation between exon density and genetic diversity could be expected and, albeit very weak, was indeed found ($r^2 = 0.015$, $P < 0.0001$; Fig. 3B). This may reflect that adaptation processes in beech affect quantitative, polygenically encoded traits [58], and therefore molecular signatures of selection differ only slightly from neutral expectations [57, 59, 60].

Flow cytometric genome size and GC-content estimation

The measured 2C value was 1.191±0.003 pg and the GC content was 37.34%. The between-day variation caused by random instrument drift and/or nonidentical sample preparation did not exceed 0.6%. The GC content and 2C value are in the range of previously reported estimates for *F. sylvatica* (36.7%–39.9%, 1.11–1.30

Table 1: Statistics of the completeness of *de novo* genome assembly of *Fagus sylvatica* assessed with CEGMA and BUSCO

Genome	BUSCO complete (in %)	BUSCO duplicated (in %)	BUSCO fragmented (in %)	BUSCO missing (in %)	CEGMA complete (in %)	CEGMA partial (in %)	Reference
Fagus sylvatica v1.2	94	19	1.7	3.6	82	94	This study
Castanea mollisima v 1.1	91	13	4.2	4.0	77	94	[19]
Quercus robur v1.0	92	10	2.7	4.8	81	96	[17]
Quercus lobata v3.0	94	11	2.4	3.0	83	98	[62]
Olea europaea v6.0	87	19	5.2	7.6	90	96	[30]
Populus trichocarpa v3.0	96	17	1.4	2.1	92	97	[63]
Eucalyptus grandis	94	5	1.8	4.7	93	100	[64]

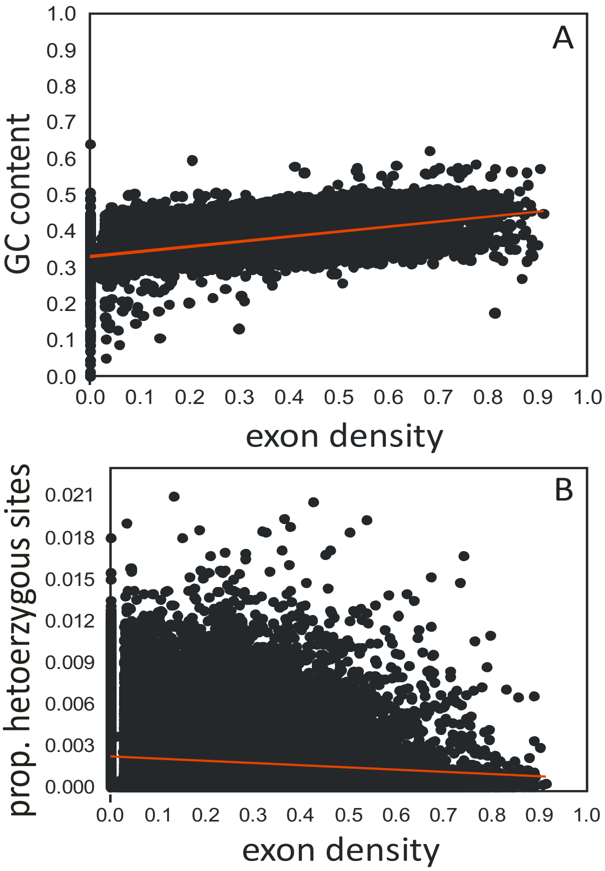


Figure 2: Parameter correlations in the *Fagus sylvatica* genome. **(A)** Gene density vs the GC content in each of the 50,994 nonoverlapping 10-kb windows. (B) Gene density vs. the proportion of heterozygous sites.

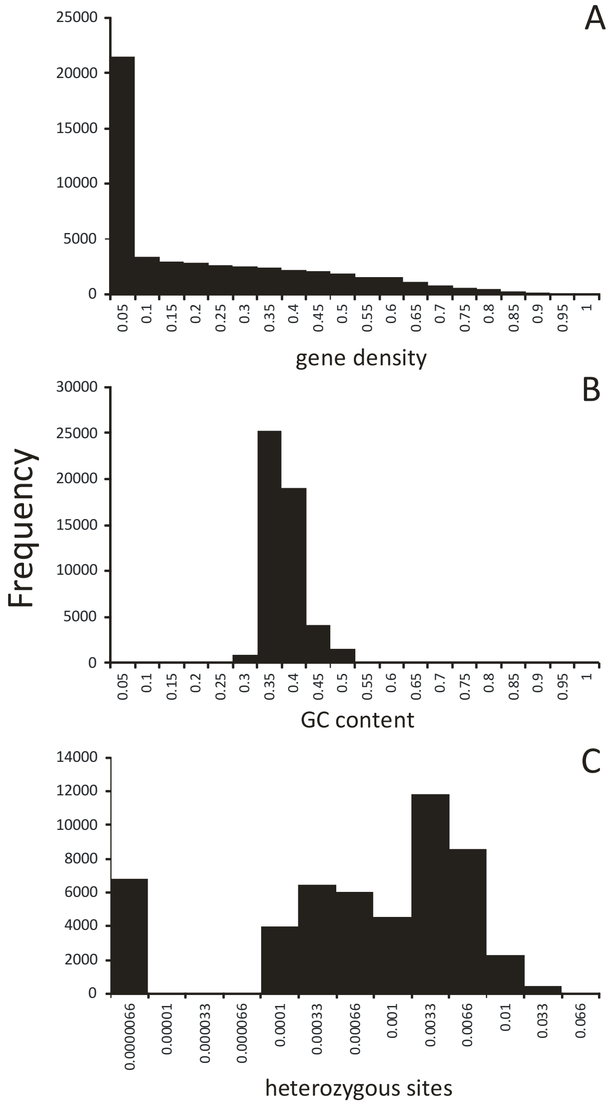


Figure 3: Parameter frequency distributions in 50-994 nonoverlapping 10-kb windows. **(A)** Gene density, measured as proportion of the window annotated as gene. **(B)** Proportion of GC bases. **(C)** Genetic diversity, measured as proportion of heterozygous sites.

pg; [55, 61]). Interestingly, when compared to the data from the European distribution of *F. sylvatica* measured from leaves using the same methodology, the studied sample matches with the geographically nearby sample from Gruenewald, Luxembourg [61].

After conversion of the 2C value to number of bases (1 pg = 978 Mb), the 1C genome was calculated to be 582.399 Mb. This value is reasonably close to the draft genome assembly. The difference of approximately 40 Mb can likely be attributed to the collapsing of centromeric and telomeric repeats in the assembly.

Genome completeness

The CEGMA analysis for evaluating assembly completeness and continuity showed a high level of completeness, with 242 of 248 (94%) of the Core Eukaryotic Genes (CEGs) at least partially covered, including 213 CEGs (82%) considered complete as per CEGMA criteria [41]. A BUSCO analysis revealed the retrieval of 94% of complete BUSCO genes, of which 19% were duplicated. Only 1.7% of the BUSCO genes were reported as fragmented and 3.6% were reported to be missing from the genome (Table 1). This places the genome among other high-quality draft genomes for tree species. In total, 75.47% of the shotgun reads used in the assembly mapped back to the assembly uniquely and in correct orientation, covering 532 Mb of the assembly.

Checks for contamination

As numerous fungi have been reported to be associated with beech [29], special attention was paid to screen for potential fungal contamination. Gene models of *F. sylvatica* were used as query in a homology-based search using BLAST against two databases, one containing the genes of *Arabidopsis thaliana* and the other containing genes from *Fungi* (both extracted from the NCBI nucleotide database), and revealed 222 genic regions with a fungal organism as the top-hit. When these 222 genes were again used as queries in a homology-based search using BLAST against the NR database from NCBI, eight genes were resolved as still having fungal top hits. These eight genes were manually inspected for the distribution of conservation. As conservation was always below a BLAST alignment score of 200 and conserved features were short, there was no conclusive evidence to support that potential contaminant fungi have impacted the assembly. In a MEGAN analysis of the genome chopped into 300 nucleotide fragments, the fragments were either categorized into flowering plants or left unassigned, suggesting a contamination load below detection threshold.

Re-use potential

The European beech is arguably one of the most important and iconic hardwood tree species in central Europe, where it forms monospecific stands under optimal growing conditions, outcompeting all other European broad-leaved tree species. Thus, there is a keen interest in the ecological genetics and genomics of the species. With the present genomic resources and the established genome browser, we provide a solid foundation for future investigations, giving the data provided a high re-use potential. In addition, the European beech genome adds to the few tree genomes published so far and is likely to be used in a variety of comparative genomics studies. Furthermore, this data resource build based on the individual “Bhaga,” will be part of a large pan-European consortium studying the genomic adaptation of beech and will serve as the reference genome and a cornerstone for future investigations.

Availability of supporting data

Raw data and assemblies were deposited in the ENA with the project accession PRJEB24056. In addition, the genome and annotation can be accessed and browsed at www.beechgenome.net. Custom scripts, annotations, and other supporting data are also available from the *GigaScience* GigaDB repository [65].

Additional files

Figure S1. Kmer-based genome size estimation.
Figure S2. Percentage of genes plotted against the number of exons in a given gene.
Figure S3. Distribution of the minimum distance among annotated genes in base pairs.
Figure S4. Distribution of distances among heterozygous sites in base pairs.

Abbreviations

Blast: Basic Local Alignment Search Tool ; CEG:Core Eukaryotic Genes ; ENA: European Nucleotide Archive; LJD: long-jumping-distance; NCBI: National Center for Biotechnology Information; RNA-seq: RNA sequencing; SMRT: Single Molecule Real-Time

Competing interests

The authors declare that they have no competing interests.

Funding

This project was partially supported by the Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE), in the framework of the Biodiversity and Climate Research Centre (BiK-F; M.P., M.T., T.H.), Integrative Pilzforschung (IPF; M.T.), and Translationale Biodiversitätsgenomik (TBG; M.P., M.T.). J.B., B.U., and J.W. were supported by grant 2012/04/A/NZ9/00500 from the National Science Center, Poland.

Authors’ contributions

M.T. conceived the project. M.T. and B.N. collected samples, J.P. conducted experiments, and B.N. extracted genomic DNA and RNA. B.M., D.K.G., and R.S. assembled the genome, provided annotations, and set up the genome browser. B.M., B.U., D.K.G., J.W., M.P., and M.T. analyzed the genome. B.M., E.L., J.B., J.P., M.P., M.T., and T.H. wrote the manuscript, with contributions from the other authors. All authors read and approved the final manuscript.

Acknowledgements

The Kellerwald-Edersee National Park is gratefully acknowledged for allowing the sequencing of the individual Bhaga.

References

1. San-Miguel-Ayanz J, de Rigo D, Caudullo G, et al. European Atlas of Forest Tree Species. Publication Office of the European Union: Luxembourg. 2016. ISBN: 978-92-79-36740-3.
2. Ellenberg H, Leuschner C. Vegetation Mitteleuropas Mit Den Alpen, 6th Edition. Eugen Ulmer KG: Stuttgart; 2010.
3. UNESCO: UNESCO World Heritage sites. http://whc.unesco.org/en/list/ (2017). Accessed 30 March 2018.
4. Langer E, Langer G, Popa F, et al. Naturalness of selected European beech forests reflected by fungal inventories: a first checklist of fungi of the UNESCO World Natural Heritage Kellerwald-Edersee National Park in Germany. Mycol Prog 2015;**14**:102.
5. Pena R. Functional diversity of beech (*Fagus sylvatica* L.) ectomycorrhizas with respect to nitrogen nutrition in response to plant carbon supply. Cuviller Verlag: Göttingen; 2011.
6. Farr DF, Rossman AY. Fungal Databases, U.S. National Fungus Collections, ARS, USDA. https://nt.ars-grin.gov/fungaldatabases/ 2017. Accessed 18 Dec 2017.
7. Heilmann-Clausen J, Aude E, Christensen M. Cryptogam communities on decaying deciduous wood – does tree species diversity matter? Biodiv Cons 2005;**14**:2061–78.

8. Ódor P, Heilmann-Claussen J, Christensen M, et al. Diversity of dead wood inhabiting fungal and bryophyte assemblages in semi-natural beech forests in Europe. Biol Cons 2006;**131**:58–71.
9. Christensen M, Heilmann-Claussen J, Walleyn R, et al. Wood-inhabiting fungi as indicators of nature value in European beech forests. Monitoring and Indicators of Forest Biodiversity in Europe - From Ideas to Operationality. EFI Proceedings No. 51; 2004.
10. Leberecht M, Dannemann M, Gschewndtner S, et al. Ectomycorrhizal communities on the roots of two beech (*Fagus sylvatica*) populations from contrasting climates differ in nitrogen acquisition in a common environment. Appl Env Microbiol 2015;**81**:5957–67.
11. Bohn U, Neuhäusle R, Gollub G. et al. Map of the Natural Vegetation of Europe. Landwirtschaftsverlag Münster; 2003.
12. Brus D, Hengeveld G, Walvoort D. et al. Statistical mapping of tree species over Europe. Europ J Forest Res 2012;**131**:145–57
13. Gessler A, Keitel C, Kreuzwieser J, et al. Potential risks for European beech (*Fagus sylvatica* L.) in a changing climate. Trees 2007;**21**:1–11.
14. Kramer K, Degen B, Buschbom J, et al. Modelling exploration of the future of European beech (*Fagus sylvatica* L.) under climate change - Range, abundance, genetic diversity and adaptive response. Forest Ecol Manag 2010;**259**:2213–22.
15. La Porta N, Capretti P, Thomsen IM, et al. Forest pathogens with higher damage potential due to climate change in Europe. Can J Pl Pathol 2008;**30**:177–95.
16. Lindner M, Maroschek M, Netherer S, et al. Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. Forest Ecol Manag 2010;**259**:698–709.
17. Plomion C, Aury JM, Amselem J, et al. Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Mol Ecol Res 2016;**16**:254–65.
18. Sork VL, Fitz-Gibbon ST, Puiu D, et al. First draft assembly and annotation of the genome of a California endemic oak *Quercus lobata* Née (Fagaceae). G3 2016;**6**:3485–95.
19. Hardwood Genomics Project: *Castanea mollisima*. https://www.hardwoodgenomics.org/chinese-chestnut-genome. Accessed 30 March 2018.
20. Lalagüe H, Csilléry K, Oddou-Muratorio S, et al. Nucleotide diversity and linkage disequilibrium at 58 stress response and phenology candidate genes in a European beech (*Fagus sylvatica* L.) population from southeastern France. Tree Gen Genomes 2014;**10**:15–26.
21. Csilléry K, Lalagüe H, Vendramin GG, et al. Detecting short spatial scale local adaptation and epistatic selection in climate-related candidate genes in European beech (*Fagus sylvatica*) populations. Mol Ecol 2014;**23**:4696–708.
22. Müller M, Seifert S, Finkeldey R. A candidate gene-based association study reveals SNPs significantly associated with bud burst in European beech (*Fagus sylvatica* L.). Tree Gen Genomes 2015;**11**:116.
23. Krajmerová D, Hrivnák M, Ditmarová Ľ, et al. Nucleotide polymorphisms associated with climate, phenology and physiological traits in European beech (*Fagus sylvatica* L.). New Forests 2017;**48**:463–77.
24. Pluess AR, Frank A, Heiri C, et al. Genome–environment association study suggests local adaptation to climate at the regional scale in *Fagus sylvatica*. New Phytologist 2016;**210**:589–601.
25. Ćalić I, Koch J, Carey D, et al. Genome-wide association study identifies a major gene for beech bark disease resistance in American beech (*Fagus grandifolia* Ehrh.). BMC Genomics 2017;**18**:547.
26. Hrivnák M, Krajmerová D, Frýdl J, et al. Variation of cytosine methylation patterns in European beech (*Fagus sylvatica* L.). Tree Gen Genomes 2016;**13**:117.
27. Lesur I, Bechade A, Lalanne C, et al. A unigene set for European beech (*Fagus sylvatica* L.) and its use to decipher the molecular mechanisms involved in dormancy regulation. Mol Ecol Res 2015;**15**:1192–204.
28. Müller M, Seifert S, Lübbe T, et al. De novo transcriptome assembly and analysis of differential gene expression in response to drought in European beech. PloS one 2017;**12**:e0184167.
29. Unterseher M, Peršoh D, Schnittler M. Leaf-inhabiting endophytic fungi of European beech (*Fagus sylvatica* L.) co-occur in leaf litter but are rare on decaying wood of the same host. Fungal Div 2013;**60**:43–54.
30. Cruz F, Julca I, Gómez-Garrido J, et al. Genome sequence of the olive tree, *Olea europaea*. GigaScience 2016;**5**:29.
31. Ali T, Schmuker A, Runge F, et al. Morphology, phylogeny, and taxonomy of *Microthlaspi* (Brassicaceae: Coluteocarpeae) and related genera. Taxon 2016;**65**:79–98.
32. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 1987;**19**:11–15.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;**30**:2114–20.
34. Joshi NA, Fass JN. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). https://github.com/najoshi/sickle 2015. Accessed 30 March 2018.
35. Hackl T, Hedrich R, Schultz J, et al. Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 2014;**30**:3004–11.
36. EBI: European Nucleotide Archive. https://www.ebi.ac.uk/ena. Accessed 30 March 2018.
37. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;**18**:821–9.
38. Ye C, Hill CM, Wu S, et al. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep 2016;**6**:31900.
39. Boetzer M, Henkel CV, Jansen HJ, et al. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2010;**27**:578–9.
40. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;**27**:764–70.
41. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007;**23**:1061–7.
42. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**:3210–2.
43. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;**14**:R36.
44. Conesa A, Götz S, García-Gómez JM, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;**21**:3674–6.
45. NCBI: RefSeq database. ftp://ftp.ncbi.nlm.nih.gov/blast/db/. Accessed 30 March 2018.
46. Hoff J. BRAKER2. http://bioinf.uni-greifswald.de/augustus/binaries/BRAKER2.tar.gz 2017. Accessed 30 March 2018.
47. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped

RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res 2014;**42**:e119.

48. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 2003;**19**(Suppl 2):II215–25.
49. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Proceedings of the13 Annual International Conference on Intelligent Systems for Molecular Biology (ISMB-05). Detroit, Michigan, 2005.
50. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0 1996-2010; http://www.repeatmasker.org. Accessed 30 March 2018.
51. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008;**18**:1851–8.
52. NCBI: NR database ftp://ftp.ncbi.nlm.nih.gov/blast/db/ 2017. Accessed 30 March 2018.
53. Huson DH, Beier S, Flade I, et al. MEGAN community edition – interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comp Biol 2016;**12**:e1004957.
54. Mishra B, Gupta DK, Thines M. The beech genome online resource (BeGOR). http://www.beechgeneome.net 2017. Accessed 30 March 2018.
55. Gallois A, Burrus M, Brown S. Evaluation of the nuclear DNA content and GC percent in four varieties of *Fagus sylvatica* L. Ann Forest Sci 1999;**56**:615–8.
56. Glémin S, Clément Y, David J, et al. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Gen 2014;**30**: 263–70.
57. Charlesworth B. Why we are not dead one hundred times over. Evolution 2013;**67**:3354–61.
58. Gömöry D, Ditmarová Ľ, Hrivnák M, et al. Differentiation in phenological and physiological traits in European beech (*Fagus sylvatica* L.). European J Forest Res 2015;**134**:1075–85.
59. Messer PW, Ellner SP, Hairston NG. Can population genetics adapt to rapid evolution? Trends Gen 2016;**32**:408–18.
60. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nature Rev Gen 2009;**10**: 195–205.
61. Paule J, Paule L, Gömöry D. Small genome size variation across the range of European beech (*Fagus sylvatica* L.). Plant Syst Evol 2018;**304**:577.
62. Valley Oak Genome Project. *Quercus mollisima* assembly v3. https://valleyoak.ucla.edu/genomicresources 2017. Accessed 30 March 2018.
63. Tuskan GA, Difazio S, Jansson S, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 2006;**313**:1596–604.
64. Myburg AA, Grattapaglia D, Tuskan GA, et al. The genome of *Eucalyptus grandis*. Nature 2014;**510**:356–62.
65. Mishra B, Gupta DK, Pfenninger M, et al. Supporting data for "A reference genome of the European Beech (Fagus sylvatica L.)." GigaScience Database 2018. http://dx.doi.org/10.5524/100461.