bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A chromosome-level genome assembly of the European Beech (*Fagus sylvativa*) reveals anomalies for organelle DNA integration, repeat content and distribution of SNPs

Bagdevi Mishra[1,2], Bartosz Ulaszewski[3], Joanna Meger[3], Markus Pfenninger[1], Deepak K Gupta[1,2,4], Stefan Wötzel[1,2], Sebastian Ploch[1], Jaroslaw Burczyk[3], Marco Thines[1,2,4]*

[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] Kazimierz Wielki University, Department of Genetics, ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland

[4]LOEWE Centre for Translational Biodiversity Genomics (TBG), Georg-Voigt-Str. 14-16, D-60325 Frankfurt am Main (Germany)

*author for correspondence – m.thines@thines-lab.eu

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background: The European Beech is the dominant climax tree in most regions of Central Europe and valued for its ecological versatility and hardwood timber. Even though a draft genome has been published recently, higher resolution is required for studying aspects of genome architecture and recombination. **Results:** Here we present a chromosome-level assembly of the more than 300 year-old reference individual, Bhaga, from the Kellerwald-Edersee National Park (Germany). Its nuclear genome of 541 Mb was resolved into 12 chromosomes varying in length between 28 Mb and 73 Mb. Multiple nuclear insertions of parts of the chloroplast genome were observed, with one region on chromosome 11 spanning more than 2 Mb of the genome in which fragments up to 54,784 bp long and covering the whole chloroplast genome were inserted randomly. Unlike in *Arabidopsis thaliana*, ribosomal cistrons are present in *Fagus sylvatica* only in four major regions, in line with FISH studies. On most assembled chromosomes, telomeric repeats were found at both ends, while centromeric repeats were found to be scattered throughout the genome apart from their main occurrence per chromosome. The genome-wide distribution of SNPs was evaluated using a second individual from Jamy Nature Reserve (Poland). SNPs, repeat elements and duplicated genes were unevenly distributed in the genomes, with one major anomaly on chromosome 4. **Conclusions:** The genome presented here adds to the available highly resolved plant genomes and we hope it will serve as a valuable basis for future research on genome architecture and for understanding the past and future of European Beech populations in a changing climate.

Keywords – Chromosomes, Fagaceae, genome architecture, genomics, Hi-C, repeat elements, SNPs

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data Description

Background

Many lowland and mountainous forests in Central Europe are dominated by the European Beech (*Fagus sylvatica*) [1]. This tree is a shade-tolerant hardwood tree that can survive as a sapling in the understorey for decades until enough light becomes available for rapid growth and maturation [2, 3]. Beech trees reach ages of 200-300 years, but older individuals are known e.g. from suboptimal habitats, especially close to the tree line [4]. Under optimal water availability, European Beech is able to outcompete most other tree species, forming monospecific stands [5], but both stagnant soil water and drought restrict its presence in natural habitats [6, 7]. Particularly, dry summers, which have recently been observed in Central Europe and that are predicted to increase as a result of climate change [8, 9], will intensify climatic stress as already now severe damage has been observed [7, 10]. In order to cope with this, human intervention in facilitating regeneration of beech forests with more drought-resistant genotypes might be a useful strategy [11, 12]. However, for the selection of drought-resistant genotypes, whole genome sequences of trees that thrive in comparatively dry conditions and the comparison with trees that are declining in drier conditions are necessary to identify genes associated with tolerating these adverse conditions [13]. Such genome-wide association studies rely on well-assembled reference genomes onto which genome data from large-scale resequencing projects can be mapped (e.g. [14]).

Due to advances in library construction and sequencing, chromosome-level assemblies have been achieved for a variety of genomes from various kingdoms of live, including animals [15, 16, 17]. While the combination of short- and long-read sequencing has brought about a significant improvement in the assembly of the gene space and regions with moderate repeat-element presence, chromosome conformation information libraries, such as Hi-C [18], have enabled associating scaffolds across highly repetitive regions, enabling the construction of super-scaffolds of chromosomal scale (e.g. [19]). Recently, the first chromosome-level assemblies have been published for tree and shrub species, e.g.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the tea tree (*Camellia sinensis* [20]), loquat (*Eriobotrya japonica* [21]), walnut (*Juglans regia* [22]), Chinese tupelo (*Nyssa sinensis* [23]), fragrant rosewood (*Dalbergia odorifera* [24]), wheel tree (*Trochodendron aralioides* [25]), azalea (*Rhododendron simsii* [26]), agrarwood tree (*Aquilaria sinensis* [27]), and tea olive (*Osmanthus fragrans* [28]). However, such resources are currently lacking for species of the *Fagaceae*, which includes the economically and ecologically important genera *Castanea*, *Fagus*, and *Quercus* [29]. For this family, various draft assemblies have been published [30, 31, 32], including European Beech [33], but none is so far resolved on a chromosome scale. To achieve this, we have sequenced the genome of the more than 300 year-old beech individual, Bhaga, from the Kellerwald-Edersee National Park (Germany), and compared it to an individual from the Jamy Nature Reserve (Poland), to get first insights into the genome architecture and variability of *Fagus sylvatica*.

Materials and Methods

Sampling and processing

The more than 300 year-old beech individual Bhaga (Fig. 1) lives on a rocky outcrop on the edge of a cliff in the Kellerwald-Edersee National Park in Hesse, Germany (51°10'09"N 8°57'47"E). Dormant buds were collected for the extraction of high molecular weight DNA as described previously [33] and for constructing Hi-C libraries in February 2018. Hi-C libraries construction and sequencing was done by a commercial sequencing provider (BGI, Hong Kong, China). For an initial assessment of genome variability, Illumina reads derived from the Polish individual, Jamy, reported in Mishra et al. [34], were used (see below).

Chromosomal pseudo molecule building using Hi-C reads

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The previous scaffold-level assembly was constructed with Illumina shotgun short reads and PacBio long reads [33]. For a chromosome-level assembly, intermediate results from the previous assembly were used as the starting material. Sequence homology of the 6699 scaffolds generated from the DBG2OLC hybrid assembler [35] to the separately assembled chloroplast and mitochondria of Beech were inferred using blast v2.10.1 [36]. All scaffolds that match in full length to any of the Organelle with identity > 99 % and gaps and/or mismatches ≤ 3 were discarded. The remaining 6657 scaffolds along with Hi-C data (116 Mb) were used in allhic [37] for building the initial Chromosome level assembly. The cleaned Illumina reads were aligned to the initial assembly using Bowtie2 software [38] and then, sorted and indexed bam files of the concordantly aligned read pairs for all the sequences were used in Pilon [39] to improve the correctness of the assembly. The final assemblies for Bhaga and Jamy were deposited under the accession numbers PRJEB24056 and PRJNA450822, respectively.

The completeness of the assembly was evaluated with plant-specific (viridiplantae_odb10.2019-11-20) and eudicot-specific (eudicots_odb10.2019-11-20) Benchmarking Universal Single-Copy Orthologs (BUSCO v4.1.4) [40].

Gene prediction

Cleaned transcriptomic Illumina reads (minimum read length: 70; average read quality: 25 and read pairs containing no N) were aligned to the assembly using Hisat [41] in order to generate splice-aware alignments. The sorted and indexed bam file (samtools, v1.9 [42]) of the splice alignments was used in “Eukaryotic gene finding” pipeline of OmicsBox [43] which uses Augustus [44] for gene prediction. For prediction, few parameters were changed from the default values. Minimum intron length was set to 20 and minimum exon length was set to 200 and complete genes (with start and stop codon) of a minimum of 180 bp length were predicted, by choosing *Arabidopsis thaliana* as the closest organism.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Assessment of the gene space

The protein sequences of the PLAZA genes for *A. thaliana*, *Vitis vinifera*, and *Eucalyptus grandis* were downloaded from plaza v4.5 dicots [45] dataset and were used along with the predicted proteins from our assembly to make protein clusters using cd-hit v.4.8.1 [46, 47]. The number of exons per genes was assessed and compared to the complete coding genes from *A. thaliana*, *Populus trichocarpa*, and *Castanea mollissima*, in line with the comparison made in the scaffold level assembly [33].

Functional annotation of the genes

The predicted genes were translated into proteins using transeq (EMBOSS:6.6.0.0 [48]) and were queried against the non-redundant database from NCBI (downloaded on 2020-06-24) [49] using diamond (v0.9.30) software [50] to find homology of the predicted proteins to sequences of known functions. For prediction of protein family membership and the presence of functional domains and sites in the predicted proteins, Interproscan (v5.39.77) software [51] was used. Result files from both diamond and Interproscan (in Xml format) were used in the blast2go [52] module of OmicsBox and taking both homology and functional domains into consideration, the final functional annotations were assigned to the genes. The density of coding space for each 100 kb region stretch was calculated for all the Chromosomes.

Repeat prediction and analysis

A repeat element database was generated using RepeatScout (v1.0.5) [53], which was used in RepeatMasker (v4.0.5) [54] to predict repeat elements. The predicted repeat elements were further filtered on the basis of their copy numbers. Those repeats represented with at least 10 copies in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genome were retained as the final set of repeat elements of the genome. Repeat fractions per 100 kb region for each of the Chromosomes were calculated for accessing patterns of repeat distribution over the genome.

In a separate analysis, repeat elements present in *Fagus sylvatica* were identified by a combination of homology-based and de novo approaches using RepeatModeler 2.0 [55] and RepeatMasker v. 4.1.1 [56]. First, we identified and classified repetitive elements de novo and generated a library of consensus sequences using RepeatModeler 2.0 [55]. We then annotated repeats in the assembly with RepeatMasker 4.1.1 [56] using the custom repeat library generated in the previous step.

Telomeric and Centromeric repeat identification

Tandem repeat finder (TRF version 4.0.9) [57] was used with parameters 2, 7, 7, 80, 10, 50 and 500 for Match, Mismatch, Delta, PM, PI, Minscore and MaxPeriod respectively [22] and all tandem repeats with monomer length up to 500 bp were predicted. Repeat frequencies of all the monomers were plotted against the length of the monomers to identify all high-frequency repeats. As the repeats were fetched by TRF program with different start and end positions and the identical repeats were falsely identified as different ones, the program MARS [58] was used to align the monomers of the different predicted repeats, and the repeat frequencies were adjusted accordingly. The chromosomal locations of telomeric and centromeric repeats were identified by blasting the repeats to the chromosomes. For confirmation of centromeric locations, pericentromeres of *A. thaliana* were blasted against the chromosomes of Bhaga.

Organelle integration

Separately assembled chloroplast and mitochondrial genomes were aligned to the genomic assembly using blastn with an e-value cut-off of 10e-10. Information for different match lengths and different

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

identity cut-offs were tabulated and analysed. Locations of integration into the nuclear genome were inferred at different length cut-offs for sequence homology (identity) equal to or more than 95%. The number of insertions per non-overlapping window of 100 kb was calculated separately for both organelles.

SNP identification and assessment

The DNA isolated from the Polish individual Jamy individual was shipped to Macrogen Inc. (Seoul, Rep. of Korea) for library preparation with 350 bp targeted insert size using TruSeq DNA PCR Free preparation kit (Illumina, USA) and sequencing on HiSeq X device (Illumina, USA) using PE-150 mode. The generated 366,127,860 raw read pairs (55.3 Gb) were processed with AfterQC v 0.9.1 [59] for quality control, filtering, trimming and error removal with default parameters resulting in 54.12 Gbp of high quality data. Illumina shotgun genomic data from Jamy was mapped to the Chromosomes level assembly using stringent parameters (--very-sensitive mode of mapping) in bowtie2 [38]. The sam formatted output of Bowtie2 was converted to binary format and sorted according to the coordinates using samtools version 1.9 [42]. SNPs were called from the sorted mapped data using bcftools (version: 1.10.2) [60] call function. SNPs were called for only those genomic locations with sequencing depth ≥ 10 bases. All locations 3 bp upstream and downstream of gaps were excluded. For determining heterozygous and homozygous states in Bhaga, sites with more than one base called and a ratio between the alternate and the reference allele of ≥ 0.25 and < 0.75 in were considered as heterozygous SNP. Where the ratio was ≥ 0.75, the position was considered homozygous. In addition, homozygous SNPs were called by comparison to Jamy, where the consensus base in Jamy has different than in Bhaga and Bhaga was homozygous at that position. SNP density was calculated for each chromosome in 100 kb intervals.

Genome browser

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A genome browser was set up using JBrowse v.1.16.10 [61]. Tracks for the predicted gene model, annotated repeat elements were added using the gff files. Separate tracks for the SNP locations and the locations of telomere and centromere were added as bed files. A track depicting the GC content was also added. The genome browser can be accessed from http://beechgenome.net.

Results

General genome features

The final assembly of the Bhaga genome was based on hybrid assembly of PacBio and Illumina reads as well as scaffolding using a Hi-C library. It was resolved into 12 chromosomes, spanning 535.4 Mb of the genome and 155 unassigned contigs of 4.9 Mb, which to 79% consisted of unplaced repeat regions that precluded their unequivocal placement. It revealed a high level of BUSCO gene detection (97.4%), surpassing that of the previous assembly and other genome assemblies available for members of the *Fagaceae* (Table 1). Of the complete assembly, 57.12% were annotated as interspersed repeat regions and 1.97% consisted of simple sequence repeats (see Supplementary File 1 for details regarding the repeat types and abundances).

The gene prediction pipeline yielded 63,736 complete genes with start and stop codon and a minimum length of 180 bp. Out of these, 2,472 genes had alternate splice variants. For 86.8% of all genes, a functional annotation could be assigned. Gene density varied widely in the genome, ranging from zero per 100 kb window to 49.7%, with an average and median of 18.2% and 17.6%, respectively. Gene lengths ranged from 180 to 54,183 bp, with an average and median gene length of 3,919 and 3,082 bp, respectively. In *Fagus sylvatica* 4.9 exons per gene were found on average, corresponding well to other high-quality plant genome drafts. The distribution of exons and introns in comparison to *J. regia* and *A. thaliana* are presented in Table 2. An analysis of PLAZA genes identified 28,326 such genes in *F. sylvatica*, out of which 1,776 genes were present in three other species used for comparison (Supplementary File 2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Telomere and centromere predictions

The tandem repeat element TTTAGGG was the most abundant repeat in the genome and was the building block of the telomeric repeats. Out of 12 chromosomes, 8 have stretches of telomeric repeats towards both ends of the chromosomes and the other 4 chromosomes have telomeric repeats towards only one end of chromosomes (Fig. 2). One unplaced scaffold of 110,653 bp which is composed of 12,051 bp of telomeric repeats at one end, probably represents one of the missing chromosome ends.

Two different types of potential centromeric repeats were observed, consisting of 79 bp and 80 bp monomer units (Supplementary File 3). Centromeric repeats were also observed in higher numbers outside the main centromeric region on several chromosomes (Supplementary File 3). However, except for chromosome 10, there was a clear clustering of centromeric repeats within each of the chromosomes, likely corresponding to the actual centromere of the respective chromosomes, and supported also by complementary evidence, such as similarities to centromeric regions of *A. thaliana*, high gypsy element content and low GC content (Supplementary File 3).

Integration of organelle DNA in the nuclear genome

For both chloroplast and mitochondria, multiple integrations of fragments of variable length of their genomic DNA were observed in all chromosomes (Figs. 3, 4). These fragments varied in length from the minimum size threshold (100 bp) to 54,784 bp for the chloroplast and 26,510 bp for the mitochondrial DNA. The identity of the integrated organelle DNA with the corresponding stretches in the organelle genome ranged from the minimum threshold tested of 95% to 100%. Nuclear-integrated fragments of organelle DNA exceeding 10 kbp were found on six chromosomes for the chloroplast, but only on one chromosome for the mitochondrial genome (Figs. 3, 4).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The integration of organelle DNA into the nuclear genome was mostly even, but tandem-like integrations of chloroplast DNA on chromosome 2 were observed (Fig. 3). In addition, insertions of both organelles were found close to the ends in 4 of the 24 chromosome ends (4, 6, 7, and 8). For the insertions further than 500 kb away from the chromosome ends the integration sites of mitochondrion DNA were sometimes found within the same 100 kb windows where the chloroplast DNA insertion was found. If some regions of the genome are more amenable for the integration of organelle DNA than others needs to be clarified in future studies. A major anomaly was found on Chromosome 11, where in a stretch of about 2 Mb consisting mainly of multiple insertions of both chloroplast and mitochondrial DNA was observed. In this region, an insertion of more than 20 kb of mitochondrial DNA was flanked by multiple very long integrations of parts of the chloroplast genome on both sides (Figs. 3, 4).

Nuclear insertions with sequence identity > 99% were about ten times more frequent for chloroplast than for mitochondrial DNA with 173 vs. 16 for fragments > 1 kb and 115 vs. 11 for fragments > 5 kb, respectively. Eight of these matches of mitochondria were located on unplaced contigs. Overall, mitochondrial insertions tended to be smaller and show a slightly higher sequence similarity (Supplementary File 4), suggesting that they might be purged from the nuclear genome quicker than the chloroplast genome insertions.

Repeat elements and gene space

The most abundant repeat elements were LTR elements and LINEs, covering 11.49% and 3.66% of the genome, respectively. A detailed list of the element types found, their abundance and proportional coverage of the genome is given in Supplementary File 1. Repeat elements presence was variable across the chromosomes (Fig. 5). While the repeat content per 100 kb window exceeded 50 % over more than 88% of chromosome 1, this was the case for only 37.5% of chromosome 9. Chromosomes showed an accumulation of repeat elements towards their ends,

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

except for chromosome 10, where only a moderate increase was observed on one of the ends, and chromosome 1, where repeat elements were more evenly distributed. Repeat content was unevenly distributed, with a patchy distribution of repeat-rich and repeat-poor regions of variable length.

A conspicuous anomaly was noticed in chromosome 4, where at one end a large region of about 10 Mb was found in which 97% of the 100 kb windows had a repeat content greater than 70%. This region also contained a high proportion of duplicated or multiplicated genes (Fig. 5). Additional regions containing more than 20% of duplicated genes within a window of at least 1 Mb were identified on chromosomes 4, 10, and 11. On chromosome 11, two clusters were detected, one of which corresponded to the site of organelle DNA insertions described above.

The ribosomal cistrons were reported to be located at the telomeres of four different chromosomes in *F. sylvatica* [58]. Due to the highly repetitive nature of the ribosomal repeats and their placement near the telomers, they could not be assigned with certainty to specific chromosomes and thus remained in four unplaced contigs. However, the 5S unit, which is separate from the other ribosomal units in *F. sylvatica*, could be placed near the centromeric locations of chromosomes 1 and 2, in line with the locations inferred by fluorescence microscopy [62].

Coding space was more evenly distributed over the chromosomes, with the exception of the regions with high levels of duplicated or multiplied genes. Apart from this, a randomly fluctuating proportion of coding space was observed, with only few regions that seemed to be slightly enriched or depleted in terms of coding space, e.g. in the central part of chromosome 8.

Distribution of single nucleotide polymorphisms

A total of 2,787,807 SNPs were identified out of which 1,271,410 SNPs were homozygous (i.e. an alternating base on both chromosomes between Bhaga and Jamy) and 1,582,804 were heterozygous

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(representing two alleles within Bhaga). A total of 269,756 SNPs fell inside coding regions out of which 119,946 were homozygous.

Heterozygous SNPs were very unequally distributed over the chromosomes (Fig. 6). Several regions, the longest of which comprised more than 30 Mb on chromosome 6, contained only very low amounts of heterozygous SNPs. Apart from the chromosome ends, where generally few heterozygous positions were observed, all chromosomes contained at least one window of 1 Mb where only very few heterozygous SNPs were present. On chromosomes 2, 3, 4, 6, and 9 such areas extended beyond 5 Mb. On chromosome 4 this region corresponded to the repeat region anomaly reported in the previous paragraph, but for the region poor in heterozygous SNPs on chromosome 9, no association with a repeat-rich region could be observed.

Homozygous SNPs differentiating Bhaga and Jamy, often followed a different pattern. All regions with low heterozygous SNP frequency longer than 5 Mb had an above-average homozygous SNP frequency, with the exception of the anomalous repeat-rich region on Chromosome 4, which had very low frequencies for both homozygous and heterozygous SNPs. However, there were also two regions of more than 1 Mb length on Chromosome 11 that also showed low frequencies of both SNP categories (Fig. 6).

Generally, the frequency of overall and intergenic SNPs per 100 kb window corresponded well for both heterozygous and homozygous SNPs, suggesting neutral evolution. However, there were some regions in which genic and intergenic SNP frequencies were uncoupled. For example, on chromosome 1 a high overall heterozygous SNP frequency was observed at 37.7, 48.2 and 56 Mb, but genic heterozygous SNP frequency was low despite normal gene density, suggesting the presence of highly conserved genes. In line with this, also the frequency of homozygous genic SNPs was equally low in the corresponding areas Similary, homozygous SNP frequencies were also decoupled on chromosome 1, where a low frequency was observed at 4.2, 7.1, 38.2, 62.1, and 64.8 Mb, but a high

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genic SNP frequency was observed. This suggests the presence of diversifying genes in the corresponding 100 kb windows, such as genes involved in coping with biotic or abiotic stress.

In line with the different distribution over the chromosomes, with large areas poor in heterozygous SNPs, there were much more windows with low numbers of heterozygous SNPs than windows with homozygous SNPs (Fig. 7). Notably, at intermediate SNP frequencies, homozygous SNPs were found in more 100 kb windows, while at very high SNP frequencies, heterozygous SNPs were more commonly found. This pattern is consistent with predominant local pollination, but occasional introgression of highly distinct genotypes.

The genome browser is available at beechgenome.net. Predicted genes, annotated repeat elements and homozygous and heterozygous SNPs are available in “B. Annotations”. The telomeric and centromeric locations and the GC content details are available in “C. Other Details”.

Discussion

General genome features

The genome assembled and analysed in this study compares well with previously published *Fagaceae* genomes, both in terms of size and gene space. We here confirm the base chromosome number of 12, as was previously reported based on chromosome counts [62]. The number of exons per gene is moderately higher than in the previously published genome of the same individual [33], reflecting the higher contiguity of the presented chromosome-level assembly. Despite the lower chromosome number of the Beech genome, it is structurally similar to the available genomes of genus *Juglans*, which is the most closely related genus for which chromosome-level assemblies are available (*J. regia* [22]; *J. sigillata* [63]; *J. regia* × *J. microcarpa* [64]).

Telomere and centromere predictions

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Telomeres are inherently difficult to resolve because of long stretches of GC-rich repeats that can cause artefacts during library preparation [65] and can lead to biased mapping [66]. However, using long-read sequencing and Hi-C scaffolding, we could identify telomeric repeats on all chromosomes. It seems likely that several of the unplaced contigs of 4.9 Mb, which included telomeric sequences, were not correctly anchored in the assembly due to ambiguous Hi-C association data resulting from the high sequence similarity of telomeric repeats, because of which for four chromosomes we could identify telomeric repeats only on one of the ends. This might also be due to the presence of ribosomal cistrons on four chromosome ends, which might have interfered with the Hi-C linkage due to their length and very high sequence similarity. On the outermost regions of the chromosomes, no longer telomeric repeat stretches were present most likely due to their ambiguous placement in the assembly, because of very high sequence similarity.

Centromere repeats were identified by screening the genome for repeats of intermediate sizes, and were found to be present predominantly within a single location per chromosome. However, lower amounts of centromeric repeat units were also observed to be scattered throughout the genome. The function of the centromeric repeats outside of the centromere remains largely enigmatic but could be associated with chromosome structuring [67] or centromere repositioning [69, 69]. Interestingly, we could find two major groups of potential centromeric repeat units of different lengths, which did not always coincide. The location of the main occurrence of the centromere-defining repeat unit agreed well with the location previously inferred using chromosome preparations and fluorescence microscopy [62].

Integration of organelle DNA in the nuclear genome

Organelle DNA integration has been frequently found in all kingdoms of life for which high-resolution genomes are available [70-72]. It can be assumed that this transfer of organelle DNA to the nucleus is the seed of transfer of chloroplast genes to the nuclear genome [73]. However, apart from a few

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hints [74] it is unclear, which factors stabilise the chloroplast genome so that its content in non-parasitic plants stays relatively stable over long evolutionary timescales [75, 76]. In the present study, it has been found that the insertion of organelle DNA insertions are located mainly in repeat-rich regions of the Beech genome. However, their presence in regions without pronounced repeat density might suggest that repeats are not the only factor associated with the insertion of organelle DNA. Nevertheless, it appears that some regions are generally amenable to the integration of organelle DNA, as in several cases chloroplast and mitochondrion insertions were observed in close proximity. The reason for this is unclear, but is known that open chromatin is more likely to accumulate insertions [77]. The potential presence of areas in the genome that are less protected from the insertion of foreign DNA could open up potential molecular biology applications for creating stable transformants.

An anomaly regarding organelle DNA insertion was observed on chromosome 11. Around a central insertion of mitochondrion DNA, multiple insertions of chloroplast DNA were found. The whole region spans more than 2 Mb, which is significantly longer than the organelle integration hotspots reported in other species [70]. The evolutionary origin of this large chromosome region is unclear, but given its repetitive nature it is conceivable that it resulted from a combination of an integration of long fragments and repeat element activity.

Distribution of single nucleotide polymorphisms (SNPs)

SNP content was found to vary across all chromosomes leading to a mosaic pattern. While most of the areas of high or low SNP density were rather short and not correlated to any other patterns, there were several regions > 1 Mbp that exhibited a similar polymorphism type, suggesting non-neutral evolution.

The longest of those stretches poor in both heterozygous and homozygous positions was found on chromosome 4, and corresponded to a region rich in both genes and repeat elements. This is

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

remarkable and probably due to a recent proliferation, as repeat-rich regions are usually less stable and more prone to accumulate mutations [78-80].

Most regions with lower abundance of heterozygous SNPs than on average were found to be particularly high in homozygous SNPs. The longest of such stretches was found on chromosome 6, comprising about two thirds of the entire chromosome. Three more such regions longer than 5 Mbp were found on other chromosomes. The evolutionary significance of this is unclear, but it is conceivable that these areas contain locale specific variants for which no alternative alleles are shared within the same stand. For confirmation of this hypothesis, it would be important to evaluate genetic markers from additional individuals of the same stand. Locally adaptive alleles could be fixed relatively easy by local inbreeding [81], considering the low seed dispersion kernel of European Beech [82]. The presence of genes involved in local adaptation could explain the rather high amount of homozygous SNPs in the same location, as the stands from which the two studied individuals came from differ in soil, water availability, continentality, and light availability. However, more individuals from geographically separated similar stands need to be investigated to disentangle the effects of inbreeding and local adaptation.

In summary, homozygous and heterozygous SNPs were rather uniformly distributed throughout the major part of the genome, suggesting neutral evolution or balancing selection.

Conclusions

The chromosome-level assembly of the ultra-centennial individual Bhaga from the Kellerwald-Edersee National Park in Germany and its comparison with the individual Jamy from the Jamy Nature Reserve in Poland has revealed several notable genomic features. The prediction of the telomeres and centromeres as well as ribosomal DNA corresponded well with data gained from chromosome imaging [62], suggesting state-of-the-art accuracy of the assembly. Interestingly, several anomalies were observed in the genome, corresponding to regions with abundant integrations of organelle

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DNA, low frequency of both heterozygous and homozygous SNPs, and long chromosome stretches almost homozygous but with a high frequency of SNPs differentiating the individuals.

Taken together, the data presented here suggest a strongly partitioned genome architecture and potentially divergent selection regimes in the stands of the two individuals investigated here. Future comparisons of additional genomes to the reference will help understanding the significance of variant sites identified in this study and shed light on the fundamental processes involved in local adaptation of a long-lived tree species exposed to a changing climate.

Availability of Supporting Data and Materials

The data sets supporting the results of this article are available in the GenBank repository, under the accession number PRJEB24056 for the *Fagus sylvatica* reference individual Bhaga and under the accession number PRJNA450822 for the individual Jamy.

Additional Files

Supplementary file 1. Details of annotated repeat elements in Fagus sylvatica.

Supplementary file 2. Venn diagram showing shared PLAZA proteins of *Arabidopsis thaliana* (27615), *Eucalyptus grandis* (36331), and *Vitis vinifera* (26346) with those of *Fagus sylvatica* (28326).

Supplementary file 3. Centromeric feature annotation.

Supplementary file 4. Details of the conservation of organelle DNA insertions in the nuclear genome.

Competing Interests

The authors declare that they have no competing interest.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Funding

This study was supported by grants of the German Science Foundation (Th1632-18-1), National Science Centre, Poland (2012/04/A/NZ9/00500), the Polish Ministry of Science and Higher Education under the program "Regional Initiative of Excellence" in 2019–2022 (Grant No. 008/RID/2018/19), and the LOEWE initiative of the government of Hessen in the framework of the LOEWE Centre for Translational Biodiversity Genomics (TBG).

Authors' Contributions

M.T. conceived the study. B.U., J.B., J.M., M.T., and S.P. provided materials. B.U., J.M., and S.P., conducted laboratory experiments. B.M., B.U., J.B., J.M., and M.T. analysed the data. B.M., B.U., J.B., J.M., M.P., M.T, and S.W. interpreted the data. B.M. and M.T. wrote the manuscript with contributions from the other authors. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to the rangers of the Kellerwald-Edersee National Park for providing access to the reference individual Bhaga and the permission to sample.

References

[1] Durrant TH, De Rigo D, Caudullo G. Fagus sylvatica in Europe: distribution, habitat, usage and threats. In: San-Miguel-Ayanz J, de Rigo D, Caudullo G, Durrant TH, Mauri A, editors. European atlas of forest tree species. Luxembourg: Publication Office of the European Union; 2016, pp 94–5.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[2] Wagner S, Collet C, Madsen P, Nakashizuka T, Nyland RD, Sagheb-Talebi K. Beech regeneration research: from ecological to silvicultural aspects. Forest Ecol Manag. 2010;**259**(11):2172–82.

[3] Ligot G, Balandier P, Fayolle A, Lejeune P, Claessens H. Height competition between *Quercus petraea* and *Fagus sylvatica* natural regeneration in mixed and uneven-aged stands. Forest Ecol Manag. 2013;**304**:391–8.

[4] Di Filippo A, Biondi F, Maugeri M, Schirone B, & Piovesan G. Bioclimate and growth history affect beech lifespan in the Italian Alps and Apennines. Glob Change Biol. 2012;**18**(3):960–72.

[5] Leuschner C, Meier IC, Hertel D. On the niche breadth of *Fagus sylvatica*: soil nutrient status in 50 Central European beech stands on a broad range of bedrock types. Ann For Sci 2006;**63**(4):355–68.

[6] Jump AS, Hunt JM, & Penuelas J. Rapid climate change-related growth decline at the southern range edge of *Fagus sylvatica*. Glob Change Biol 2006;**12**(11):2163–74.

[7] Geßler A, Keitel C, Kreuzwieser J, Matyssek R, Seiler W, Rennenberg H. Potential risks for European beech (*Fagus sylvatica* L.) in a changing climate. Trees 2007;**21**(1):1–11.

[8] Coumou D, Rahmstorf S. A decade of weather extremes. Nat Clim Change 2012;**2**(7):491–6.

[9] Spinoni J, Naumann G, Vogt J, Barbosa P. European drought climatologies and trends based on a multi-indicator approach. Global Planet Change 2015;**127**:50–7.

[10] Albert REIF, Xystrakis F, Gaertner S, Sayer U. Floristic change at the drought limit of European beech (*Fagus sylvatica* L.) to downy oak (Quercus pubescens) forest in the temperate climate of central Europe. Not Bot Horti Agrobo 2017;**45**(2):646–54.

[11] Rose L, Leuschner C, Köckemann B, Buschmann H. Are marginal beech (*Fagus sylvatica* L.) provenances a source for drought tolerant ecotypes? Eur J For Res 2009;**128**(4):335–43.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[12] Bolte A, Degen B. Forest adaptation to climate change - options and limitations. Landbauforsch Volk 2010;**60**(3):111–7.

[13] Pfenninger M, Reuss F, Kiebler A, Schönnenbeck P, Caliendo C, Gerber S, et al. Genomic basis of drought resistance in *Fagus sylvatica*. bioRxiv 2020; doi: 10.1101/2020.12.04.411264.

[14] Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature. 2010;**465**(7298):627–31.

[15] Michael TP, VanBuren R. Building near-complete plant genomes. Curr Opin Plant Biol 2020;**54**:26–33.

[16] Priest SJ, Yadav V, Heitman J. Advances in understanding the evolution of fungal genome architecture. F1000Research 2020;9(Faculty Rev):776.

[17] Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. bioRxiv 2020; doi: 10.1101/2020.05.22.110833.

[18] Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;**326**(5950):289–93.

[19] Yin X, Arias-Pérez A, Kitapci TH, Hedgecock D. High-Density Linkage Maps Based on Genotyping-by-Sequencing (GBS) Confirm a Chromosome-Level Genome Assembly and Reveal Variation in Recombination Rate for the Pacific Oyster *Crassostrea gigas*. G3 - Genes Genom Genet 2020;**10**(12):4691–705.

[20] Chen JD, Zheng C, Ma JQ, Jiang CK, Ercisli S, Yao MZ, et al. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. Hortic Res 2020;**7**(63):1–11.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[21] Jiang S, An H, Xu F, Zhang X. Chromosome-level genome assembly and annotation of the loquat (Eriobotrya japonica) genome. GigaScience 2020;**9**(3):giaa015.

[22] Marrano A, Britton M, Zaini PA, Zimin AV, Workman RE, Puiu D, et al. High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome. GigaScience 2020;**9**(5):giaa050.

[23] Yang X, Kang M, Yang Y, Xiong H, Wang M, Zhang Z, et al. A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*. Sci Data 2019;**6**(282):1–7.

[24] Hong Z, Li J, Liu X, Lian J, Zhang N, Yang Z, et al. The chromosome-level draft genome of *Dalbergia odorifera*. GigaScience 2020;**9**(8): giaa084.

[25] Strijk JS, Hinsinger DD, Zhang F, Cao K. *Trochodendron aralioides*, the first chromosome-level draft genome in *Trochodendrales* and a valuable resource for basal eudicot research. GigaScience 2019;**8**(11): giz136.

[26] Yang FS, Nie S, Liu H, Shi TL, Tian XC, Zhou SS, et al. Chromosome-level genome assembly of a parent species of widely cultivated azaleas. Nat Commun 2020;**11**(1):1–13.

[27] Nong W, Law ST, Wong AY, Baril T, Swale T, Chu LM, et al. Chromosomal-level reference genome of the incense tree *Aquilaria sinensis*. Mol Ecol Resour 2020;**20**(4):971.

[28] Yang X, Yue Y, Li H, Ding W, Chen G, Shi T, et al. The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus fragrans*. Hortic Res 2018;**5**(72):1–13.

[29] Kremer A, Abbott AG, Carlson JE, Manos PS, Plomion C, Sisco P, et al. Genomics of *Fagaceae*. Tree Genet Genomes 2012;**8**(3):583–610.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[30] Sork VL, Squire K, Gugger PF, Steele SE, Levy ED, Eckert AJ. Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*. American J Bot 2016;**103**(1):33–46.

[31] Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, et al. The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. Plant J 2016;**87**(5):507–32.

[32] Plomion C, Aury JM, Amselem J, Alaeitabar T, Barbe V, Belser C, et al. Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Mol Ecol Resour 2016;**16**(1):254–65.

[33] Mishra B, Gupta DK, Pfenninger M, Hickler T, Langer E, Nam B, et al. A reference genome of the European beech (*Fagus sylvatica* L.). GigaScience 2018;**7**(6):giy063.

[34] Mishra B, Ulaszewski B, Ploch S, Burczyk J, & Thines M. A Circular Chloroplast Genome of Fagus sylvatica Reveals High Conservation between Two Individuals from Germany and One Individual from Poland and an Alternate Direction of the Small Single-Copy Region. Forests. 2021;**12**(2):180.

[35] Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep 2016;**6**(1):1–9.

[36] Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;**215**(3):403–10.

[37] Zhang X, Zhang S, Zhao Q Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants 2019;**5**(8):833–45.

[38] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357–9.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[39] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS ONE 2014;**9**(11):e112963.

[40] Seppey M, Manni M, & Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, editors. Gene Prediction. Methods in Molecular Biology, vol 1962. New York: Humana. 2019. pp 227–45.

[41] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;**12**(4):357–60.

[42] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinf 2009;**25**(16):2078–2079.

[43] OmicsBox - Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019, https://www.biobam.com/omicsbox

[44] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 2005;**33**(suppl_2):W465–7.

[45] PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics Nucleic Acids Res (online access). Accessed 21st October, 2020.

[46] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinf 2006;**22**(13):1658–9.

[47] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;**28**(23):3150–2.

[48] Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet 2000;**16**(6):276–7.

[49] NCBI nr database https://ftp.ncbi.nlm.nih.gov/blast/db/ accessed on 24th of June 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[50] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Meth. 2015;**12**:59–60.

[51] Jones P, Binns D, Chang H, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinf 2014;**30**(9):1236–40.

[52] Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 2008;**36**(10):3420–35.

[53] Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinf 2005;**21**(suppl_1):i351–8.

[54] Smit AFA, Hubley R. RepeatMasker Open-4.0.5. 2007–2014; http://www.repeatmasker.org. Accessed 16 Nov 2020.

[55] Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. PNAS 2020;**117**(17):9451–7.

[56] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Prot Bioinf 2009;**25**(1):4.10.1–4.10.14.

[57] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;**27**(2):573–80.

[58] Ayad LA, Pissis SP. MARS: improving multiple circular sequence alignment using refined sequences. BMC Genomics 2017;**18**(86):1–10.

[59] Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinf 2017;**18**(3):80. doi:10.1186/s12859-017-1469-3

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[60] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinf 2011;**27**(21):2987–93.

[61] *Buels* R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 2016;**17**(66)1–12.

[62] Ribeiro T, Loureiro J, Santos C, Morais-Cecílio L. Evolution of rDNA FISH patterns in the *Fagaceae*. Tree Genet & Genomes. 2011;**7**(6):1113-22.

[63] Ning DL, Wu T, Xiao LJ, Ma T, Fang WL, Dong RQ, & Cao FL. Chromosomal-level assembly of *Juglans sigillata* genome using Nanopore, BioNano, and Hi-C analysis. GigaScience 2020;**9**(2):giaa006.

[64] Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, et al. Sequencing a *Juglans regia* × *J. microcarpa* hybrid yields high-quality genome assemblies of parental species. Hortic Res 2019;**6**(55):1–16.

[65] Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 2011;**12**(R18):1–14.

[66] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008;**36**(16):e105.

[67] Alves S, Ribeiro T, Inácio V, Rocheta M, Morais-Cecílio L. Genomic organization and dynamics of repetitive DNA sequences in representatives of three *Fagaceae* genera. Genome. 2012;**55**(5):348–59.

[68] Mandáková T, Hloušková P, Koch MA, Lysak MA. Genome evolution in *Arabideae* was marked by frequent centromere repositioning. Plant Cell 2020;**32**(3):650–65.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[69] Klein SJ, O'Neill RJ. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Res 2018;**26**(1):5–23.

[70] Zhang GJ, Dong R, Lan LN, Li SF, Gao WJ, Niu HX. Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. Int J Mol Sci 2020;**21**(3):707.

[71] Guo X, Ruan S, Hu W, Cai D, Fan L. Chloroplast DNA insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved. Funct Integr Genomic 2008;**8**(2):101–8.

[72] Stegemann S, Hartmann S, Ruf S, Bock R. High-frequency gene transfer from the chloroplast genome to the nucleus. PNAS 2003;**100**(15):8828–33.

[73] Huang CY, Ayliffe MA, & Timmis JN. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 2003;**422**(6927):72–6.

[74] Yang Z, Hou Q, Cheng L, Xu W, Hong Y, Li S, et al. RNase H1 cooperates with DNA gyrases to restrict R-loops and maintain genome integrity in *Arabidopsis* chloroplasts. Plant Cell 2017;**29**(10):2478–97.

[75] Xiong AS, Peng RH, Zhuang J, Gao F, Zhu B, Fu XY, et al. Gene duplication, transfer, and evolution in the chloroplast genome. Biotechnol Adv 2009;**27**(4):340–7.

[76] Wang D, Wu YW, Shih ACC, Wu CS, Wang YN, & Chaw SM. Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA. Mol Biol Evol 2007;**24**(9):2040–8.

[77] Wang D, & Timmis JN. Cytoplasmic organelle DNA preferentially inserts into open chromatin. Genome Biol Evol 2013;**5**(6):1060–4.

[78] Wang L, Sun Y, Sun X, Yu L, Xue L, He Z, et al. Repeat-induced point mutation in *Neurospora crassa* causes the highest known mutation rate and mutational burden of any cellular life. Genome Biol 2020;**21**(142):1–23.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[79] Flynn JM, Lower SE, Barbash DA, Clark AG. Rates and patterns of mutation in tandem repetitive DNA in six independent lineages of Chlamydomonas reinhardtii. Genome Biol Evol 2018;**10**(7):1673–86.

[80] Ho EK, Bellis ES, Calkins J, Adrion JR, Latta LC, Schaack S. Engines of change: Transposable element mutation rates are high and vary widely among genotypes and populations of Daphnia magna. bioRxiv 2020; doi: 10.1101/2020.09.21.307181.

[81] Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nat Rev Gen 2018;**19**(4):220.

[82] Martínez I, González-Taboada F. Seed dispersal patterns in a temperate forest during a mast event: performance of alternative dispersal kernels. Oecologia 2009;**159**(2):389–400.

[83] Wang J, Tian S, Sun X, Cheng X, Duan N, Tao J, Shen G. Construction of Pseudomolecules for the Chinese Chestnut (*Castanea mollissima*) Genome. G3 2020;**10**(10):3565–74.

Tables

Table 1. Comparison of BUSCO completeness in Fagaceae genomes available and in the present study (*Fagus sylvatica* V2).

Species	Complete	Single	Duplicated	Fragmented	Missing
Fagus sylvatica V2	97.4%	90.3%	7.1%	1.3%	1.3%
Fagus sylvatica V1 [33]	96.6%	85.6%	11%	1.8%	1.6%
Castanea mollissima [83]	92.4%	88.8%	3.7%	1.5%	6.1%

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Quercus lobata [30] v3	93.5%	87.6%	5.9%	1.0%	5.5%

Table 2. Distribution of exons in *Fagus sylvatica* in comparison to *Juglans regia* and *Arabidopsis thaliana.*

Species	Minimum exons / gene	First quartile	Mean exons / gene	Median exons / gene	Third quartile	Maximum exons / gene
Fagus sylvatica v2	1	2	4.916	4	7	70
Juglans regia [31]	1	2	5.301	4	7	70
Arabidopsis thaliana [GCA_000001735]	1	1	5.299	4	7	79

Figure captions

Fig. 1. The more than 300 year-old *Fagus sylvatica* reference individual Bhaga on a cliff over the Edersee in the Kellerwald Edersee National Park (Germany)

Fig. 2. Locations of probable centromeric repeats on the chromosomes presented as red lines and telomeric locations as blue line on the chromosomes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3. Chloroplast genome insertions within 100 kb windows on the chromosomes. Each chromosome is represented as three rows, the first with insertions more than 100 bp long, the second row with more than 1 kb and the third with more than 10 kb.

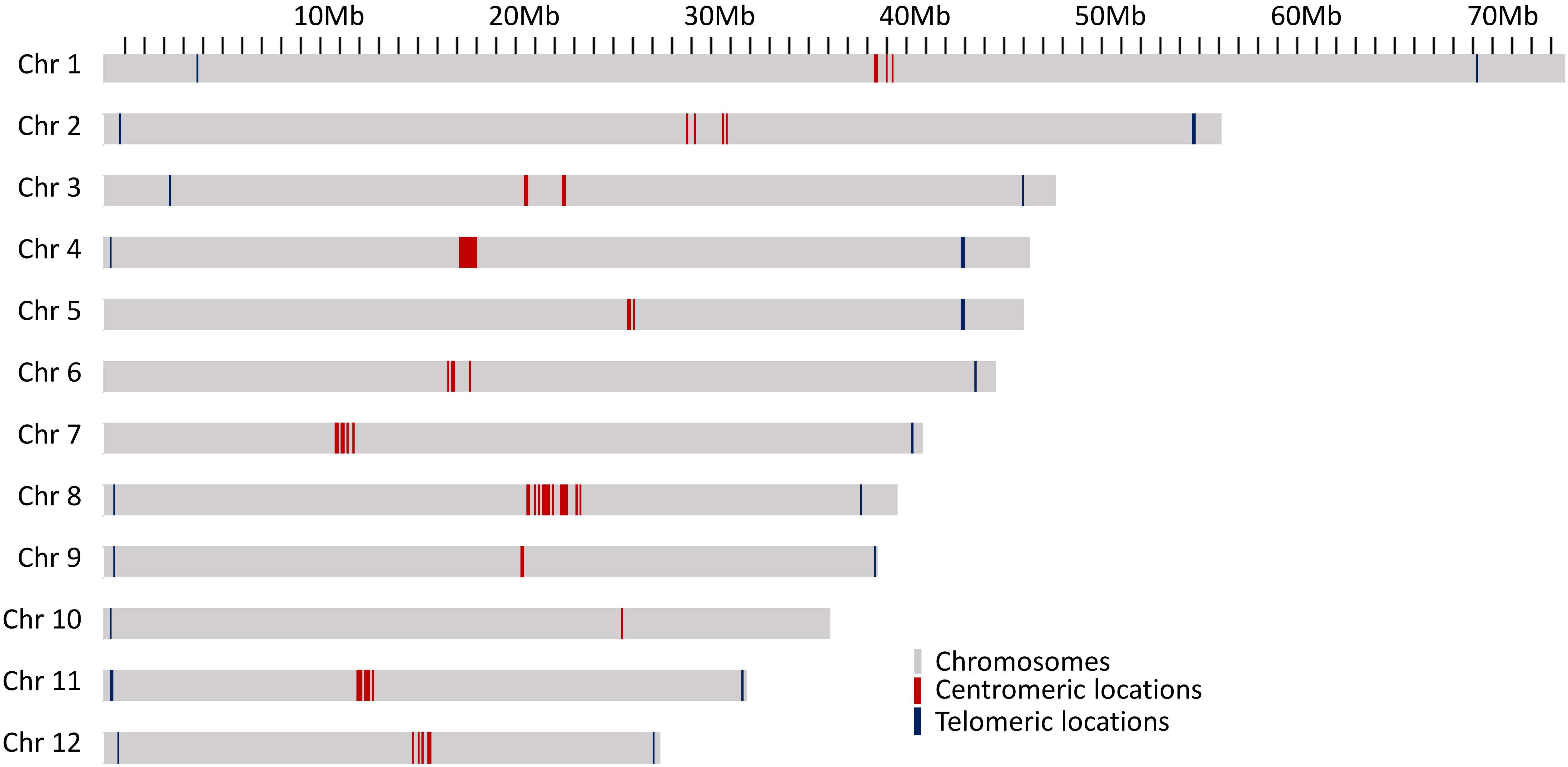
Fig. 4. Mitochondrion genome insertions within 100 kb windows on the chromosomes. Each chromosome is represented as three rows, the first with insertions more than 100 bp long, the second row with more than 1 kb and the third with more than 10 kb.

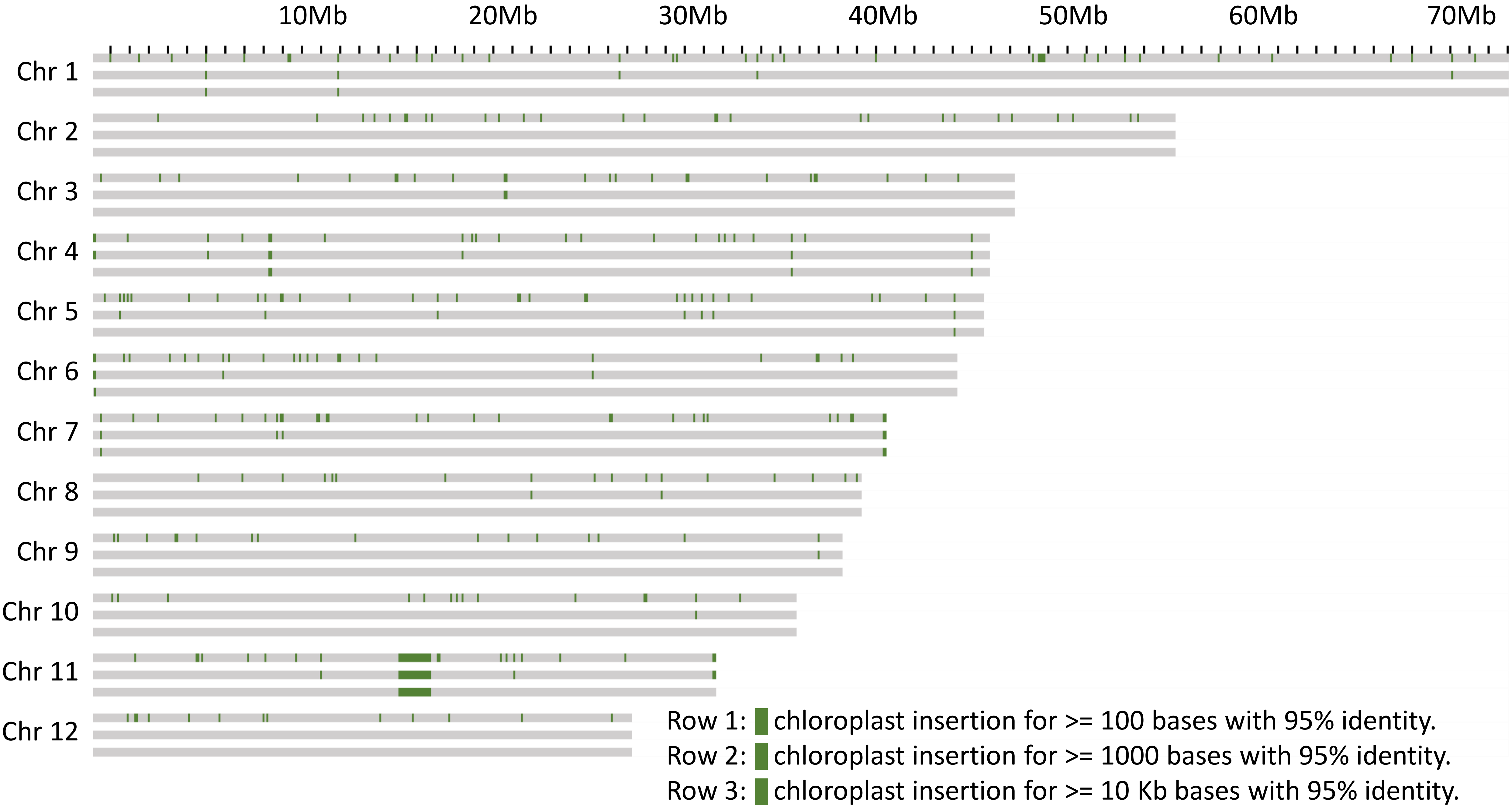
Fig. 5. Repeat regions, coding regions, and regions coding for genes present within 100 kb windows on the chromosomes.

Fig. 6. *Fagus sylvatica* Homozygous and Heterozygous SNPs present within 100 kb windows on the chromosomes.

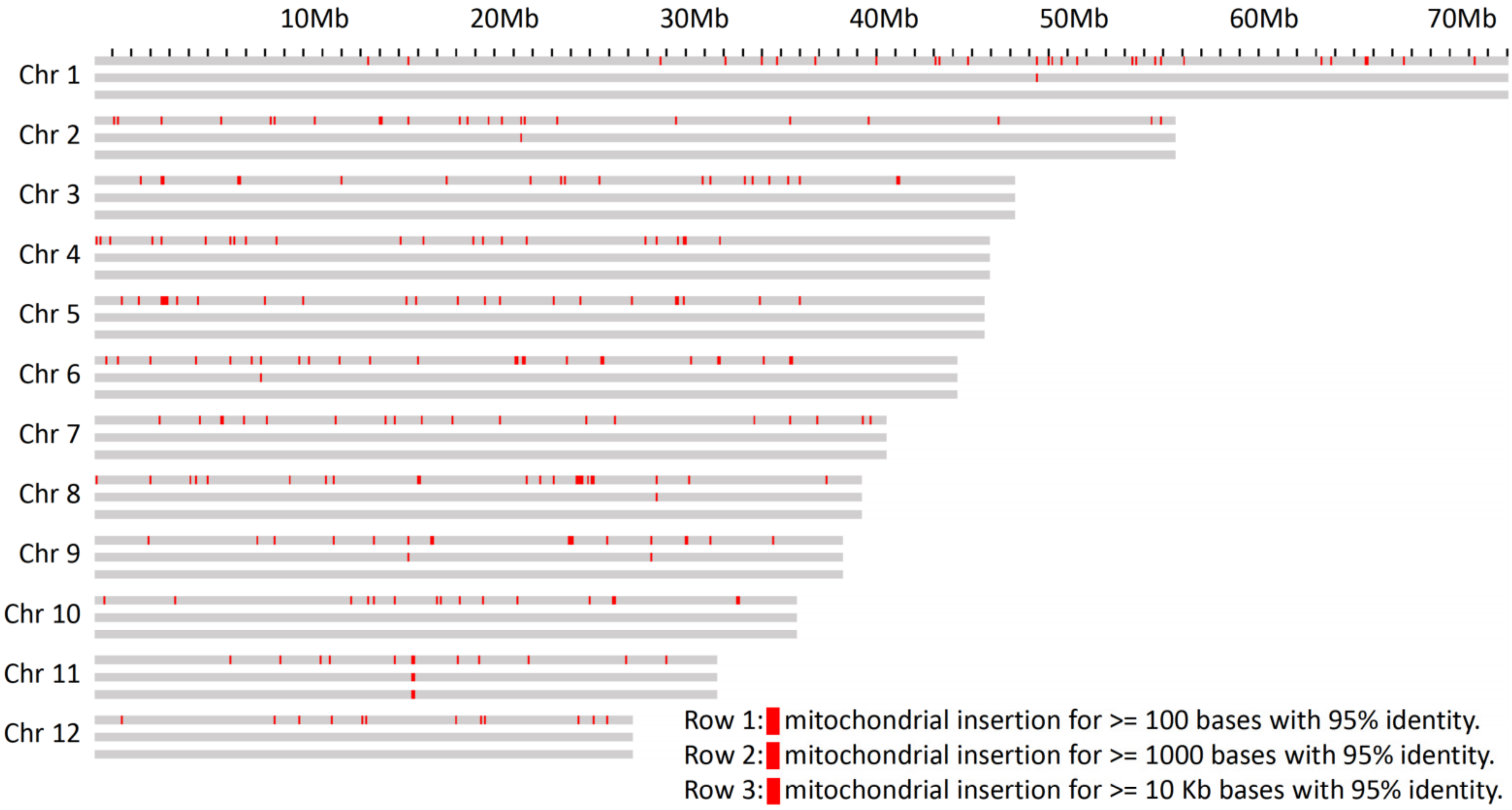
Fig. 7: Distribution of homozygous and heterozygous SNPS in non-overlapping 100 kb windows.

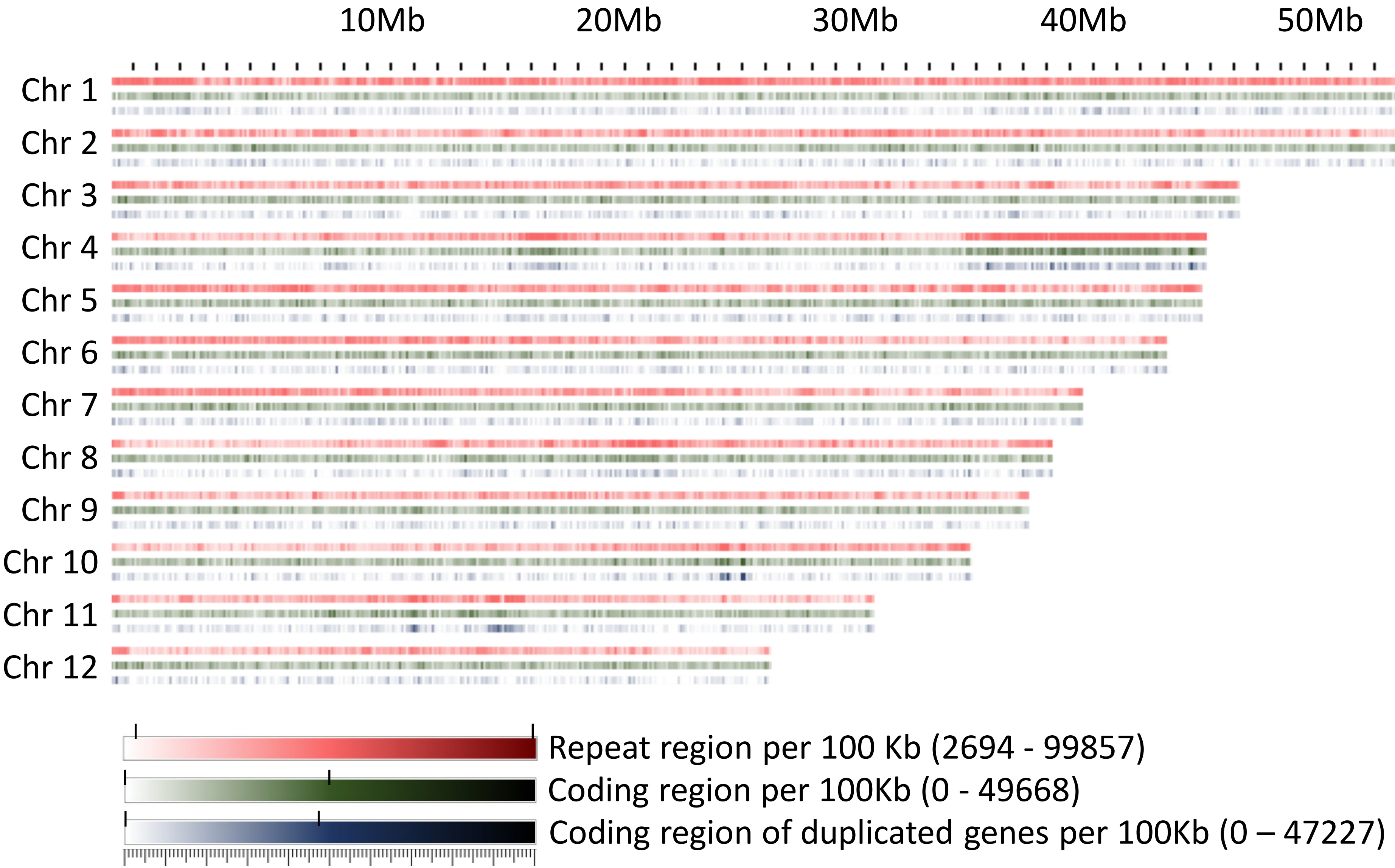
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.22.436437; this version posted March 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

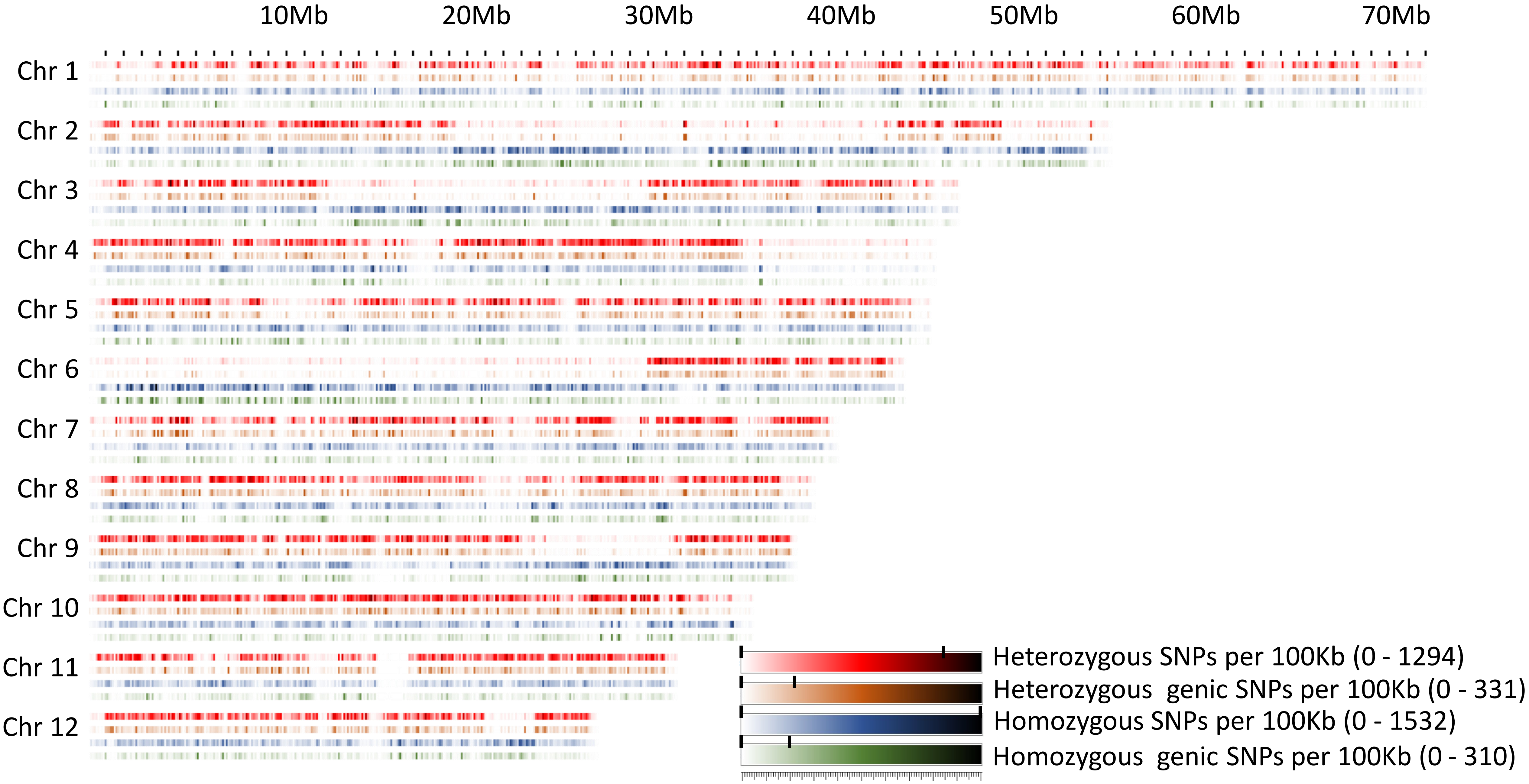

10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chromosomes
Centromeric locations
Telomeric locations

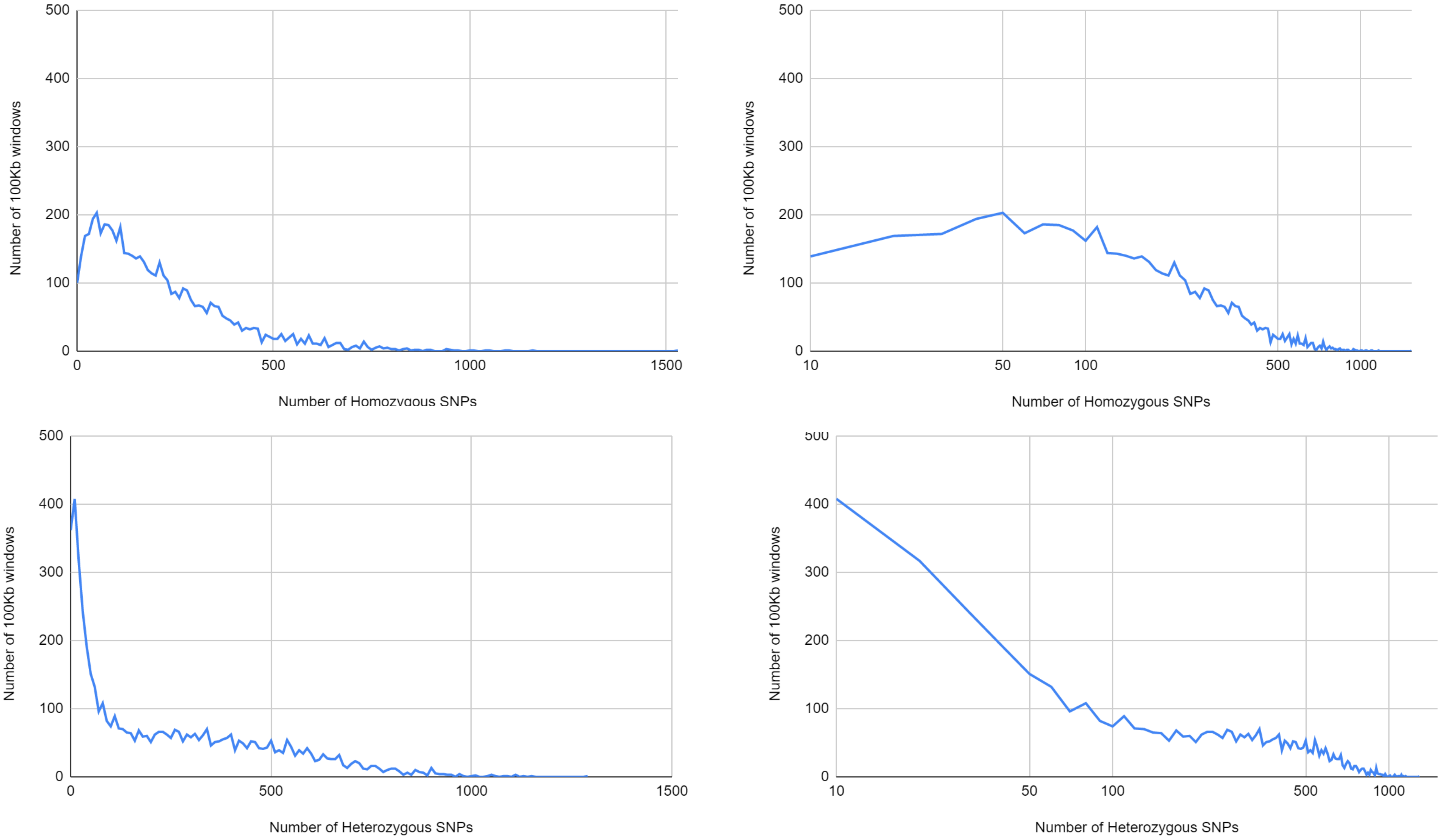


10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Row 1: 🟩 chloroplast insertion for >= 100 bases with 95% identity.
Row 2: 🟩 chloroplast insertion for >= 1000 bases with 95% identity.
Row 3: 🟩 chloroplast insertion for >= 10 Kb bases with 95% identity.


10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Row 1: 🟥 mitochondrial insertion for >= 100 bases with 95% identity.
Row 2: 🟥 mitochondrial insertion for >= 1000 bases with 95% identity.
Row 3: 🟥 mitochondrial insertion for >= 10 Kb bases with 95% identity.


10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Repeat region per 100 Kb (2694 - 99857)
Coding region per 100Kb (0 - 49668)
Coding region of duplicated genes per 100Kb (0 – 47227)


10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Heterozygous SNPs per 100Kb (0 - 1294)
Heterozygous genic SNPs per 100Kb (0 - 331)
Homozygous SNPs per 100Kb (0 - 1532)
Homozygous genic SNPs per 100Kb (0 - 310)



500
400
300
200
100
0
Number of 100Kb windows
0
500
1000
1500
Number of Homozygous SNPs
500
400
300
200
100
0
Number of 100Kb windows
10
50
100
500
1000
Number of Homozygous SNPs
500
400
300
200
100
0
Number of 100Kb windows
0
500
1000
1500
Number of Heterozygous SNPs
500
400
300
200
100
0
Number of 100Kb windows
10
50
100
500
1000
Number of Heterozygous SNPs