A chromosome-level genome assembly of the European Beech (*Fagus sylvatica*) 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

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

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.

70 the tea tree (*Camellia sinensis* [20]), loquat (*Eriobotrya japonica* [21]), walnut (*Juglans regia* [22]),
71 Chinese tupelo (*Nyssa sinensis* [23]), fragrant rosewood (*Dalbergia odorifera* [24]), wheel tree
72 (*Trochodendron aralioides* [25]), azalea (*Rhododendron simsii* [26]), agrarwood tree (*Aquilaria*
73 *sinensis* [27]), and tea olive (*Osmanthus fragrans* [28]). However, such resources are currently lacking
74 for species of the *Fagaceae*, which includes the economically and ecologically important genera
75 *Castanea*, *Fagus*, and *Quercus* [29]. For this family, various draft assemblies have been published [30,
76 31, 32], including European Beech [33], but none is so far resolved on a chromosome scale. To
77 achieve this, we have sequenced the genome of the more than 300 year-old beech individual, Bhaga,
78 from the Kellerwald-Edersee National Park (Germany), and compared it to an individual from the
79 Jamy Nature Reserve (Poland), to get first insights into the genome architecture and variability of
80 *Fagus sylvatica*.

81

82 **Materials and Methods**

83 *Sampling and processing*

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

91

92 *Chromosomal pseudo molecule building using Hi-C reads*

93 The previous scaffold-level assembly was constructed with Illumina shotgun short reads and PacBio
94 long reads [33]. For a chromosome-level assembly, intermediate results from the previous assembly
95 were used as the starting material. Sequence homology of the 6699 scaffolds generated from the
96 DBG2OLC hybrid assembler [35] to the separately assembled chloroplast and mitochondria of Beech
97 were inferred using blast v2.10.1 [36]. All scaffolds that match in full length to any of the Organelle
98 with identity > 99 % and gaps and/or mismatches ≤ 3 were discarded. The remaining 6657 scaffolds
99 along with Hi-C data (116 Mb) were used in allhic [37] for building the initial Chromosome level
100 assembly. The cleaned Illumina reads were aligned to the initial assembly using Bowtie2 software
101 [38] and then, sorted and indexed bam files of the concordantly aligned read pairs for all the
102 sequences were used in Pilon [39] to improve the correctness of the assembly. The final assemblies
103 for Bhaga and Jamy were deposited under the accession numbers PRJEB24056 and PRJNA450822,
104 respectively.

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

108

109 *Gene prediction*

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

118

119 *Assessment of the gene space*

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

126

127 *Functional annotation of the genes*

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

137

138 *Repeat prediction and analysis*

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

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

145 In a separate analysis, repeat elements present in *Fagus sylvatica* were identified by a combination
146 of homology-based and de novo approaches using RepeatModeler 2.0 [55] and RepeatMasker v.
147 4.1.1 [56]. First, we identified and classified repetitive elements de novo and generated a library of
148 consensus sequences using RepeatModeler 2.0 [55]. We then annotated repeats in the assembly
149 with RepeatMasker 4.1.1 [56] using the custom repeat library generated in the previous step.

150

151 *Telomeric and Centromeric repeat identification*

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

162

163 *Organelle integration*

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

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

170

171 *SNP identification and assessment*

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

189

190 *Genome browser*

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

195

196 **Results**

197 *General genome features*

198 The final assembly of the Bhaga genome was based on hybrid assembly of PacBio and Illumina reads
199 as well as scaffolding using a Hi-C library. It was resolved into 12 chromosomes, spanning 535.4 Mb
200 of the genome and 155 unassigned contigs of 4.9 Mb, which to 79% consisted of unplaced repeat
201 regions that precluded their unequivocal placement. It revealed a high level of BUSCO gene detection
202 (97.4%), surpassing that of the previous assembly and other genome assemblies available for
203 members of the *Fagaceae* (Table 1). Of the complete assembly, 57.12% were annotated as
204 interspersed repeat regions and 1.97% consisted of simple sequence repeats (see Supplementary File
205 1 for details regarding the repeat types and abundances).

206 The gene prediction pipeline yielded 63,736 complete genes with start and stop codon and a
207 minimum length of 180 bp. Out of these, 2,472 genes had alternate splice variants. For 86.8% of all
208 genes, a functional annotation could be assigned. Gene density varied widely in the genome, ranging
209 from zero per 100 kb window to 49.7%, with an average and median of 18.2% and 17.6%,
210 respectively. Gene lengths ranged from 180 to 54,183 bp, with an average and median gene length of
211 3,919 and 3,082 bp, respectively. In *Fagus sylvatica* 4.9 exons per gene were found on average,
212 corresponding well to other high-quality plant genome drafts. The distribution of exons and introns
213 in comparison to *J. regia* and *A. thaliana* are presented in Table 2. An analysis of PLAZA genes
214 identified 28,326 such genes in *F. sylvatica*, out of which 1,776 genes were present in three other
215 species used for comparison (Supplementary File 2).

216

217 *Telomere and centromere predictions*

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

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

231

232 *Integration of organelle DNA in the nuclear genome*

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

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

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

257

258 *Repeat elements and gene space*

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

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

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

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

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

284

285 *Distribution of single nucleotide polymorphisms*

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

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

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

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

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

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

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

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

356

357 *Integration of organelle DNA in the nuclear genome*

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

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

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

378

379 *Distribution of single nucleotide polymorphisms (SNPs)*

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

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

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

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

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

403

404 **Conclusions**

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

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

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

418

419 **Availability of Supporting Data and Materials**

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

423

424 **Additional Files**

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

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

428 **Supplementary file 3**. Centromeric feature annotation.

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

430

431 **Competing Interests**

432 The authors declare that they have no competing interest.

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.

[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.

[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.

502 [21] Jiang S, An H, Xu F, Zhang X. Chromosome-level genome assembly and annotation of the loquat
503 (Eriobotrya japonica) genome. GigaScience 2020;**9**(3):giaa015.

504 [22] Marrano A, Britton M, Zaini PA, Zimin AV, Workman RE, Puiu D, et al. High-quality chromosome-
505 scale assembly of the walnut (*Juglans regia* L.) reference genome. GigaScience
506 2020;**9**(5):giaa050.

507 [23] Yang X, Kang M, Yang Y, Xiong H, Wang M, Zhang Z, et al. A chromosome-level genome assembly
508 of the Chinese tupelo *Nyssa sinensis*. Sci Data 2019;**6**(282):1–7.

509 [24] Hong Z, Li J, Liu X, Lian J, Zhang N, Yang Z, et al. The chromosome-level draft genome of
510 *Dalbergia odorifera*. GigaScience 2020;**9**(8): giaa084.

511 [25] Strijk JS, Hinsinger DD, Zhang F, Cao K. *Trochodendron aralioides*, the first chromosome-level
512 draft genome in *Trochodendrales* and a valuable resource for basal eudicot research.
513 GigaScience 2019;**8**(11): giz136.

514 [26] Yang FS, Nie S, Liu H, Shi TL, Tian XC, Zhou SS, et al. Chromosome-level genome assembly of a
515 parent species of widely cultivated azaleas. Nat Commun 2020;**11**(1):1–13.

516 [27] Nong W, Law ST, Wong AY, Baril T, Swale T, Chu LM, et al. Chromosomal-level reference genome
517 of the incense tree *Aquilaria sinensis*. Mol Ecol Resour 2020;**20**(4):971.

518 [28] Yang X, Yue Y, Li H, Ding W, Chen G, Shi T, et al. The chromosome-level quality genome provides
519 insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus*
520 *fragrans*. Hortic Res 2018;**5**(72):1–13.

521 [29] Kremer A, Abbott AG, Carlson JE, Manos PS, Plomion C, Sisco P, et al. Genomics of *Fagaceae*.
522 Tree Genet Genomes 2012;**8**(3):583–610.

523 [30] Sork VL, Squire K, Gugger PF, Steele SE, Levy ED, Eckert AJ. Landscape genomic analysis of
524 candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*.
525 American J Bot 2016;**103**(1):33–46.

526 [31] Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, et al. The
527 walnut (Juglans regia) genome sequence reveals diversity in genes coding for the
528 biosynthesis of non-structural polyphenols. Plant J 2016;**87**(5):507–32.

529 [32] Plomion C, Aury JM, Amselem J, Alaeitabar T, Barbe V, Belser C, et al. Decoding the oak genome:
530 public release of sequence data, assembly, annotation and publication strategies. Mol Ecol
531 Resour 2016;**16**(1):254–65.

532 [33] Mishra B, Gupta DK, Pfenninger M, Hickler T, Langer E, Nam B, et al. A reference genome of the
533 European beech (*Fagus sylvatica* L.). GigaScience 2018;**7**(6):giy063.

534 [34] Mishra B, Ulaszewski B, Ploch S, Burczyk J, & Thines M. A Circular Chloroplast Genome of Fagus
535 sylvatica Reveals High Conservation between Two Individuals from Germany and One
536 Individual from Poland and an Alternate Direction of the Small Single-Copy Region. Forests.
537 2021;**12**(2):180.

538 [35] Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: efficient assembly of large genomes using long
539 erroneous reads of the third generation sequencing technologies. Sci Rep 2016;**6**(1):1–9.

540 [36] Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ. Basic local alignment search tool. J Mol
541 Biol. 1990;**215**(3):403–10.

542 [37] Zhang X, Zhang S, Zhao Q Ming R, Tang H. Assembly of allele-aware, chromosomal-scale
543 autopolyploid genomes based on Hi-C data. Nat Plants 2019;**5**(8):833–45.

544 [38] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods
545 2012;**9**(4):357–9.

546 [39] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool
547 for comprehensive microbial variant detection and genome assembly improvement. PloS
548 ONE 2014;**9**(11):e112963.

549 [40] Seppey M, Manni M, & Zdobnov EM. BUSCO: assessing genome assembly and annotation
550 completeness. In: Kollmar M, editors. Gene Prediction. Methods in Molecular Biology, vol
551 1962. New York: Humana. 2019. pp 227–45.

552 [41] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements.
553 Nat Methods 2015;**12**(4):357–60.

554 [42] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map
555 format and SAMtools. Bioinf 2009;**25**(16):2078–2079.

556 [43] OmicsBox - Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019,
557 https://www.biobam.com/omicsbox

558 [44] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that
559 allows user-defined constraints. Nucleic Acids Res 2005;**33**(suppl_2):W465–7.

560 [45] PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics
561 Nucleic Acids Res (online access). Accessed 21st October, 2020.

562 [46] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or
563 nucleotide sequences. Bioinf 2006;**22**(13):1658–9.

564 [47] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing
565 data. Bioinformatics. 2012;**28**(23):3150–2.

566 [48] Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite.
567 Trends Genet 2000;**16**(6):276–7.

568 [49] NCBI nr database https://ftp.ncbi.nlm.nih.gov/blast/db/ accessed on 24th of June 2020.

569 [50] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Meth.
570 2015;**12**:59–60.

571 [51] Jones P, Binns D, Chang H, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale
572 protein function classification. Bioinf 2014;**30**(9):1236–40.

573 [52] Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput
574 functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res
575 2008;**36**(10):3420–35.

576 [53] Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinf
577 2005;**21**(suppl_1):i351–8.

578 [54] Smit AFA, Hubley R. RepeatMasker Open-4.0.5. 2007–2014; http://www.repeatmasker.org.
579 Accessed 16 Nov 2020.

580 [55] Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for
581 automated genomic discovery of transposable element families. PNAS 2020;**117**(17):9451–7.

582 [56] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic
583 sequences. Curr Prot Bioinf 2009;**25**(1):4.10.1–4.10.14.

584 [57] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res
585 1999;**27**(2):573–80.

586 [58] Ayad LA, Pissis SP. MARS: improving multiple circular sequence alignment using refined
587 sequences. BMC Genomics 2017;**18**(86):1–10.

588 [59] Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. AfterQC: automatic filtering, trimming, error
589 removing and quality control for fastq data. BMC Bioinf 2017;**18**(3):80. doi:10.1186/s12859-
590 017-1469-3

591 [60] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
592 population genetical parameter estimation from sequencing data. Bioinf 2011;**27**(21):2987–
593 93.

594 [61] *Buels* R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web
595 platform for genome visualization and analysis. Genome Biol 2016;**17**(66)1–12.

596 [62] Ribeiro T, Loureiro J, Santos C, Morais-Cecílio L. Evolution of rDNA FISH patterns in the *Fagaceae*.
597 Tree Genet & Genomes. 2011;**7**(6):1113-22.

598 [63] Ning DL, Wu T, Xiao LJ, Ma T, Fang WL, Dong RQ, & Cao FL. Chromosomal-level assembly of
599 *Juglans sigillata* genome using Nanopore, BioNano, and Hi-C analysis. GigaScience
600 2020;**9**(2):giaa006.

601 [64] Zhu T, Wang L, You FM, Rodriguez JC, Deal KR, Chen L, et al. Sequencing a *Juglans regia* × *J.*
602 *microcarpa* hybrid yields high-quality genome assemblies of parental species. Hortic Res
603 2019;**6**(55):1–16.

604 [65] Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR
605 amplification bias in Illumina sequencing libraries. Genome Biol 2011;**12**(R18):1–14.

606 [66] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets
607 from high-throughput DNA sequencing. Nucleic Acids Res 2008;**36**(16):e105.

608 [67] Alves S, Ribeiro T, Inácio V, Rocheta M, Morais-Cecílio L. Genomic organization and dynamics of
609 repetitive DNA sequences in representatives of three *Fagaceae* genera. Genome.
610 2012;**55**(5):348–59.

611 [68] Mandáková T, Hloušková P, Koch MA, Lysak MA. Genome evolution in *Arabideae* was marked by
612 frequent centromere repositioning. Plant Cell 2020;**32**(3):650–65.

613 [69] Klein SJ, O'Neill RJ. Transposable elements: genome innovation, chromosome diversity, and
614 centromere conflict. Chromosome Res 2018;**26**(1):5–23.

615 [70] Zhang GJ, Dong R, Lan LN, Li SF, Gao WJ, Niu HX. Nuclear integrants of organellar DNA contribute
616 to genome structure and evolution in plants. Int J Mol Sci 2020;**21**(3):707.

617 [71] Guo X, Ruan S, Hu W, Cai D, Fan L. Chloroplast DNA insertions into the nuclear genome of rice:
618 the genes, sites and ages of insertion involved. Funct Integr Genomic 2008;**8**(2):101–8.

619 [72] Stegemann S, Hartmann S, Ruf S, Bock R. High-frequency gene transfer from the chloroplast
620 genome to the nucleus. PNAS 2003;**100**(15):8828–33.

621 [73] Huang CY, Ayliffe MA, & Timmis JN. Direct measurement of the transfer rate of chloroplast DNA
622 into the nucleus. Nature 2003;**422**(6927):72–6.

623 [74] Yang Z, Hou Q, Cheng L, Xu W, Hong Y, Li S, et al. RNase H1 cooperates with DNA gyrases to
624 restrict R-loops and maintain genome integrity in *Arabidopsis* chloroplasts. Plant Cell
625 2017;**29**(10):2478–97.

626 [75] Xiong AS, Peng RH, Zhuang J, Gao F, Zhu B, Fu XY, et al. Gene duplication, transfer, and evolution
627 in the chloroplast genome. Biotechnol Adv 2009;**27**(4):340–7.

628 [76] Wang D, Wu YW, Shih ACC, Wu CS, Wang YN, & Chaw SM. Transfer of chloroplast genomic DNA
629 to mitochondrial genome occurred at least 300 MYA. Mol Biol Evol 2007;**24**(9):2040–8.

630 [77] Wang D, & Timmis JN. Cytoplasmic organelle DNA preferentially inserts into open chromatin.
631 Genome Biol Evol 2013;**5**(6):1060–4.

632 [78] Wang L, Sun Y, Sun X, Yu L, Xue L, He Z, et al. Repeat-induced point mutation in *Neurospora*
633 *crassa* causes the highest known mutation rate and mutational burden of any cellular life.
634 Genome Biol 2020;**21**(142):1–23.

[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%

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

652

653

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

655 *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

656

657

658

659 **Figure captions**

660 **Fig. 1**. The more than 300 year-old *Fagus sylvatica* reference individual Bhaga on a cliff over the

661 Edersee in the Kellerwald Edersee National Park (Germany)

662 **Fig. 2**. Locations of probable centromeric repeats on the chromosomes presented as red lines and

663 telomeric locations as blue line on the chromosomes.

664

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

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

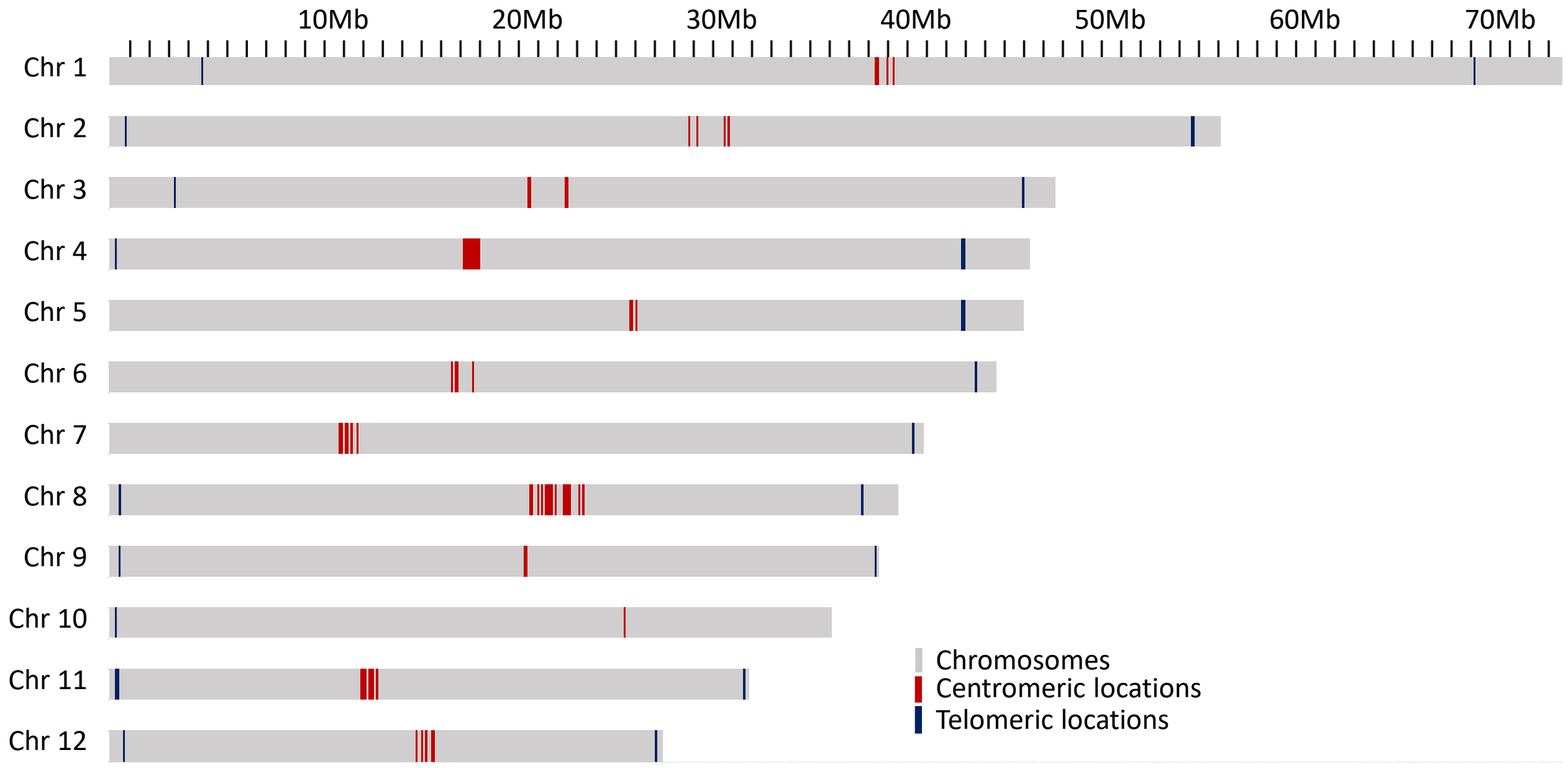
671 **Fig. 5**. Repeat regions, coding regions, and regions coding for genes present within 100 kb windows
672 on the chromosomes.

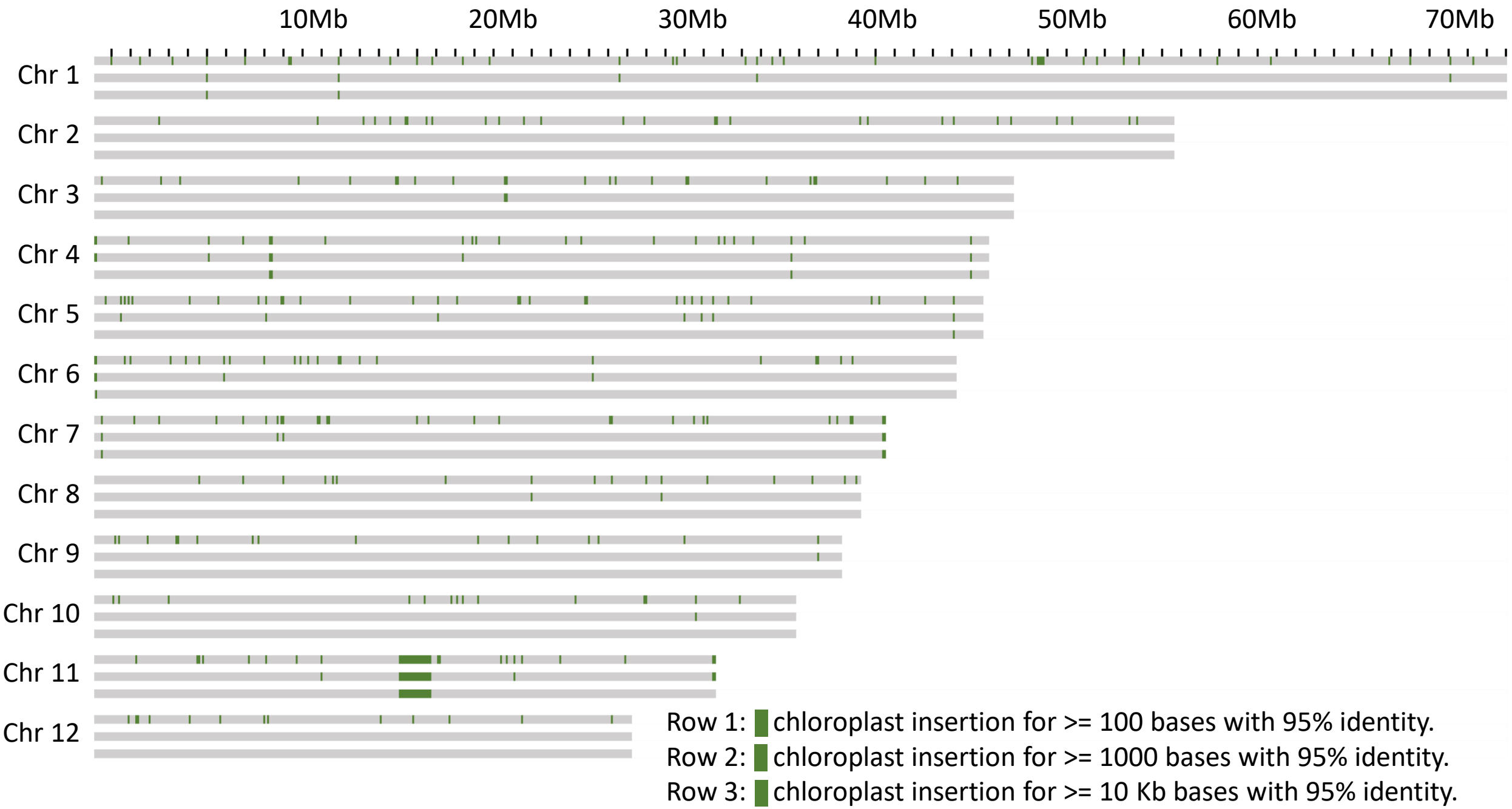
673 **Fig. 6**. *Fagus sylvatica* Homozygous and Heterozygous SNPs present within 100 kb windows on the
674 chromosomes.

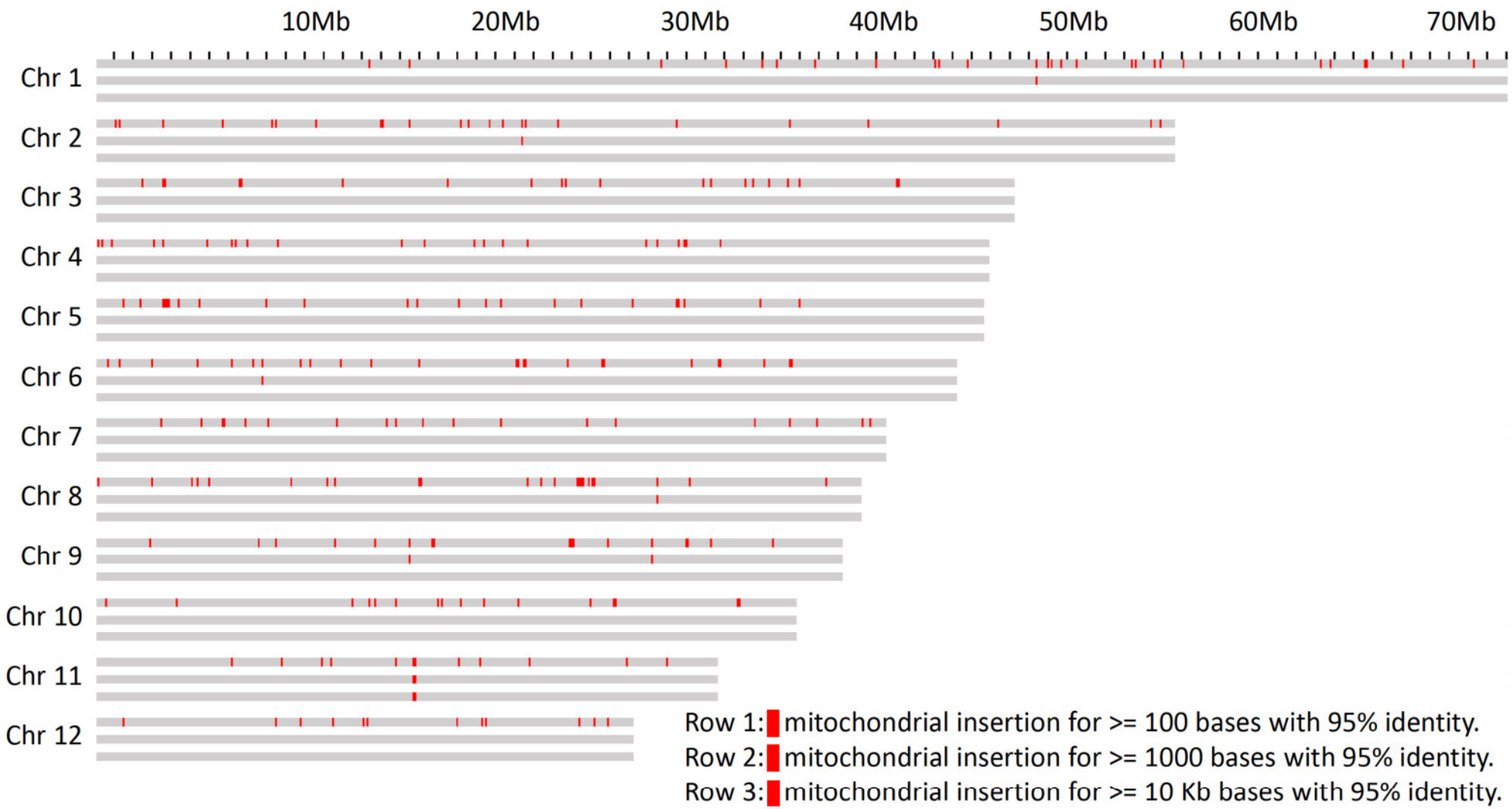
675 **Fig. 7**: Distribution of homozygous and heterozygous SNPS in non-overlapping 100 kb windows.

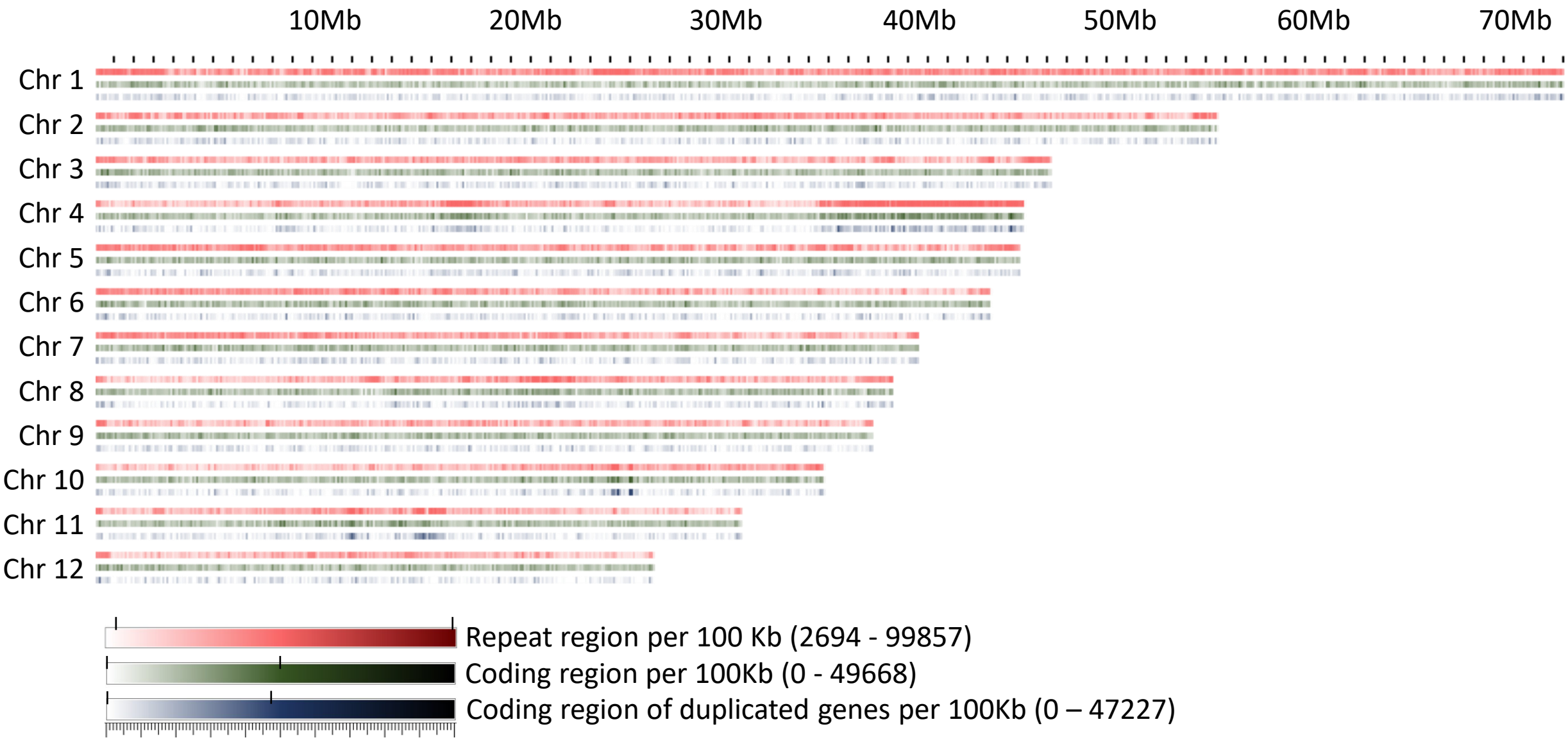
676

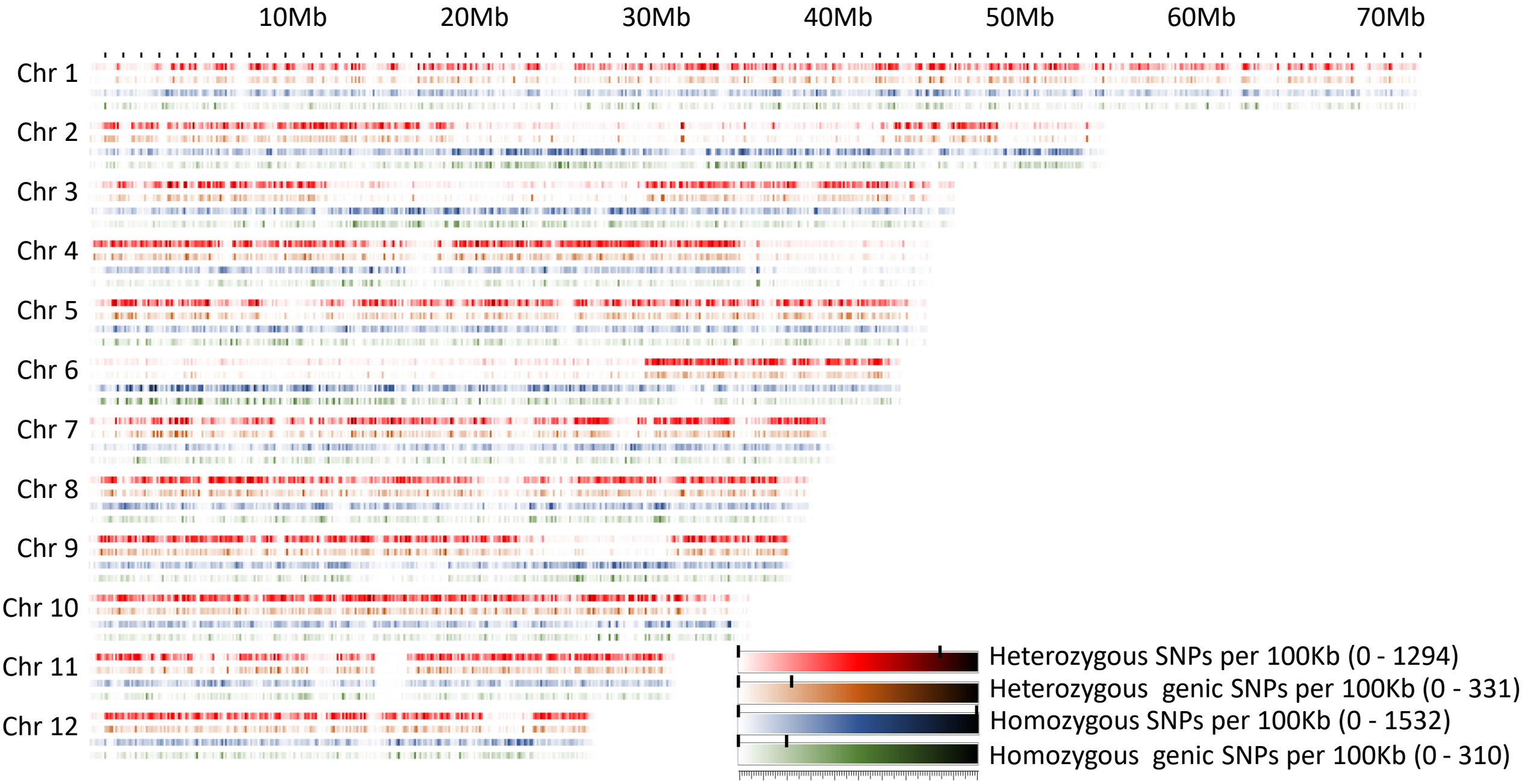
677

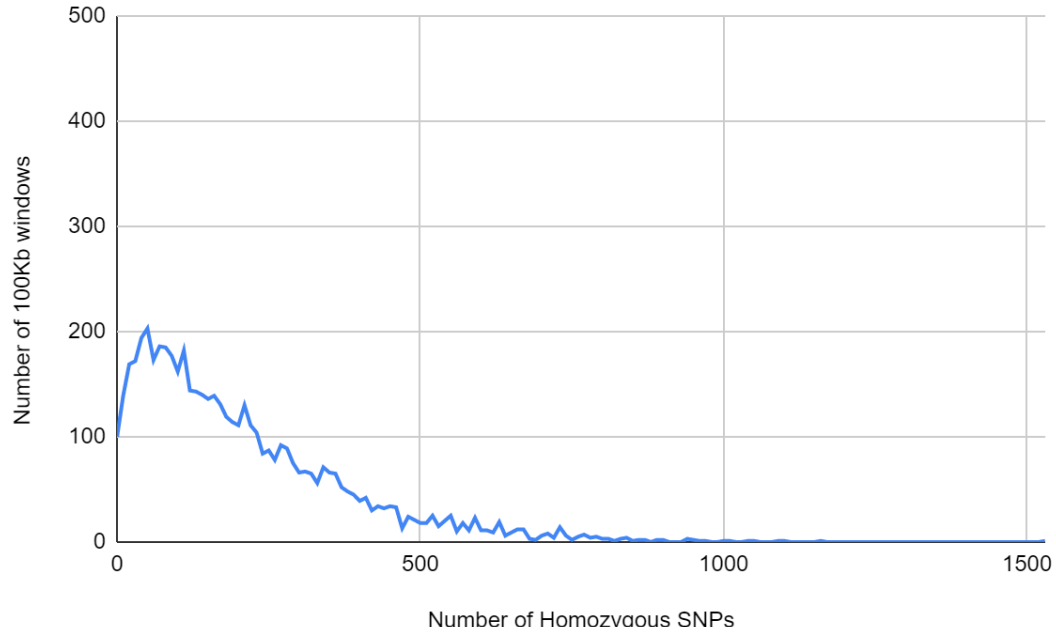

10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chromosomes
Centromeric locations
Telomeric locations

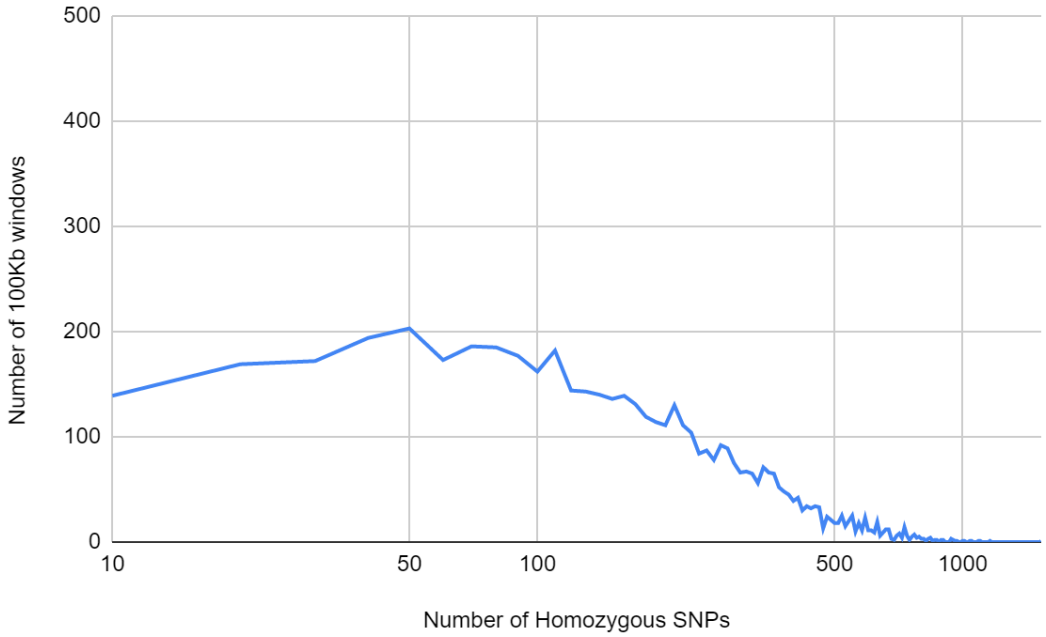

10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Row 1: chloroplast insertion for >= 100 bases with 95% identity.
Row 2: chloroplast insertion for >= 1000 bases with 95% identity.
Row 3: chloroplast insertion for >= 10 Kb bases with 95% identity.

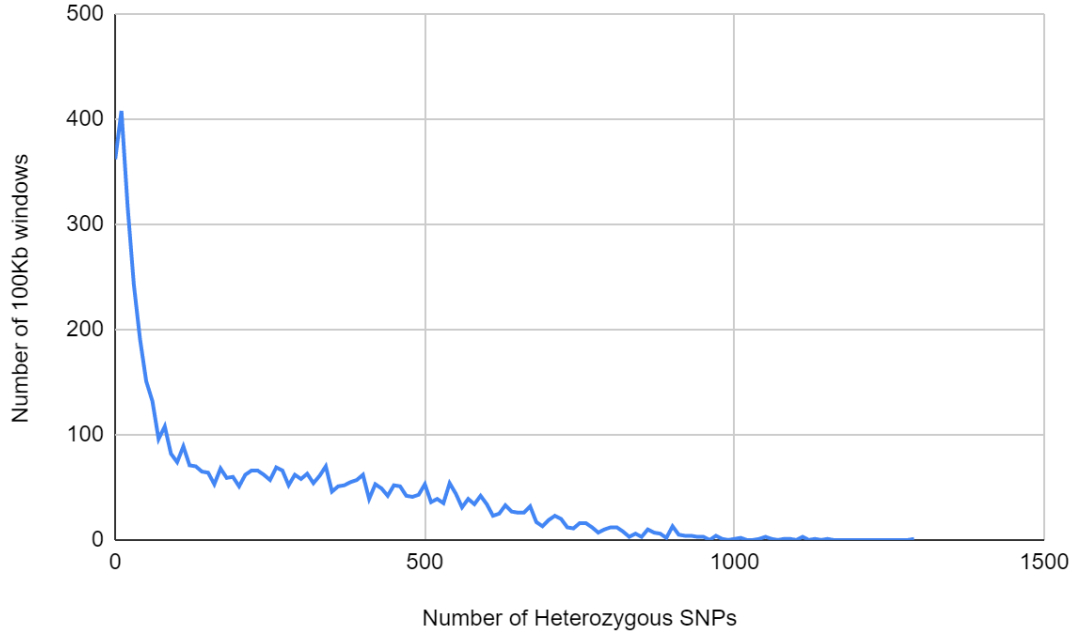

10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Row 1: mitochondrial insertion for >= 100 bases with 95% identity.
Row 2: mitochondrial insertion for >= 1000 bases with 95% identity.
Row 3: mitochondrial insertion for >= 10 Kb bases with 95%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

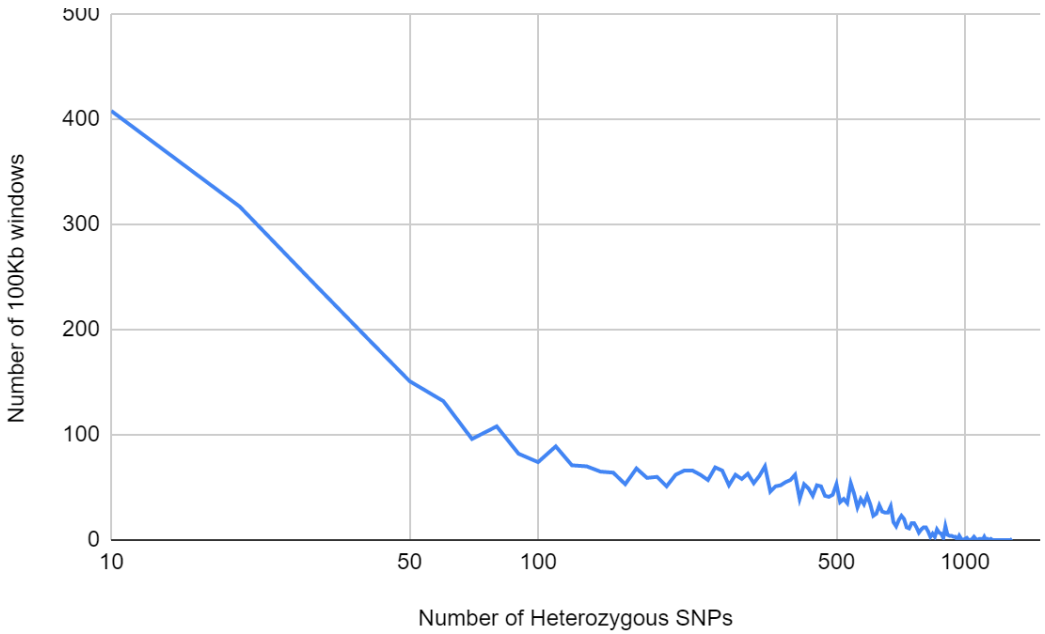

10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Repeat region per 100 Kb (2694 - 99857)
Coding region per 100Kb (0 - 49668)
Coding region of duplicated genes per 100Kb (0 – 47227)


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)


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


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


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


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