Supplementary file S1 – Additional Material & Methods, Results and Discussion

Content

1. Rationale

1.1. Using High-Throughput Sequencing (HTS) for assessing diversity in high-copy, multi-loci markers

Assuming the occurrence of at least two functional, distinct, paralogous (or hom[o]eologous) 5S rDNA arrays in the studied beech taxa (Ribeiro *et al.* 2011), our data confirm the capacity of HTS to detect both of them (see **Sections 4, 5.2**). In addition, the HTS approach captured rare sequence types pointing to past divergences (speciation events) forming part of the 5S intergenic spacer (5S-IGS) sequence pool of the sampled populations. Although the HTS results cannot be generally considered quantitative (Lamb *et al.* 2019), two main sequence types were represented in all investigated taxa (**Appendix A**; cf. main-text fig. 6). Multiple loci and polyploid genomes can severely affect the generation of comprehensive phylogenetic data covering intra-genomic variation rendering traditional PCR-based direct sequencing impossible. Special attention to sequencing enough clones to capture the signal of all existing loci is therefore crucial (e.g. Volkov *et al.* 2017), especially when recruiting ITS data (Denk *et al.* 2002; Denk, Grimm & Hemleben 2005; Grimm, Denk & Hemleben 2007a) for beech phylogenetics (four 35S rDNA loci; Ribeiro *et al.* 2011). For the 5S arrays, the required effort can be easily accomplished with HTS procedures, because of their optimal length for Miseq analyses (i.e. < 400 bp; extremely conserved flanking regions). As for our work on *Quercus* (Piredda *et al.* 2020), the 5S-IGS has demonstrated to be a very quick and efficient tool for addressing difficult evolutionary questions and ecological applications. An obvious strength of the HTS method vs traditional cloning approaches is that even low-frequent copies are captured, including copies from potentially degrading arrays. These may indicate ancient links (e.g. Grimm & Denk 2008) that would otherwise be overlooked because of ongoing concerted evolution and/or intragenomic silencing (cf. main-text fig. 8). The downside is clearly the amount of data that need to be analysed, which requires automated steps.

1.2. Sampling & pooling

So far, there are no 5S-IGS data for beeches at all. The here used samples and data represent opportunistic research. Although HTS approaches are relatively cheap next-generation sequencing methods and do not require freshly sampled material and DNA yields as high as needed for, e.g., SNP-generating next-generation sequencing approaches, financial resources and access to relatively fresh material is the most limiting factor. At the time this research was initiated the price pre sample was ~30€ plus ~4500€ for set-up and ~500€ for sequencing of up

to 192 samples. For our analysis of oaks (Piredda *et al.* 2020; work in progress), we were able to finance the HTS analysis of 250 samples in three sequencing batches. The number of samples was higher than originally planned,[1] hence, we could spare 10% for a pilot study into *Fagus* populations with a focus on the *F. sylvatica* s.str. – western *F. orientalis* hybrid zone in Greece.

However, to assess hybrids, it is necessary to first put up a data and phylogenetic framework for the focal species. First results demonstrated a nigh-complexity of the 5S-IGS pool in beeches (main-text fig. 2), surpassing the already complex situation in oak (Piredda *et al.* 2020).

Thus, for this pilot 5S-IGS study on beech, we selected four, taxonomically unambiguous, of our sequenced samples:

- One sample of a *F. sylvatica* s.str. population north of the Alps, assumed to strongly affected by Pleistocene bottlenecks;
- One sample of *F. sylvatica* s.str. south of the Alps (C. Italy), a population growing in an area with known refuges for beech during the glacials;
- One sample each representing the westernmost (NE. Greece) and easternmost *F. orientalis* (N. Iran).[2]

For comparison, we included a sample of the supposed sister species of west-Eurasian *F. sylvatica* s.l., the Japanese *F. crenata*. *Fagus crenata* is widespread in Japan and is partly sympatric with the second Japanese species *F. japonica*, a distant relative. A sample of *F. japonica* was included here as outgroup (cf. Denk, Grimm & Hemleben 2005 for inter-species phylogenetic relationships).

As for oaks (Piredda *et al.* 2020), we extracted the DNA of five individuals per population and pooled the extracts into a single sample; together with the used sequencing depths this ensures to cover both intra-genomic 5S-IGS diversity and inter-individual 5S-IGS divergence of the studied focal populations. The Japanese samples added as sister- and outgroups aimed at covering as much intra-species 5S-IGS diversity as possible, hence, we pooled individuals from five geographically distant populations (localities are included in **Fig. S3** in **Section 2.3**). While our data can be considered representative for each of the sampled western Eurasian populations, the Japanese samples should only be viewed being representative for their species across the

[1] When applying for the money, we calculated with a price of ~50€ per sample.

[2] Our sequenced beech samples unfortunately don't include any material from the Caucasus; otherwise such a sample would have been included, too.

covered range. Population-wise studies on Japanese beeches are likely to reveal further structuring of the 5S-IGS pools.

2. Dataset Description and Classification of Sequence Reads

2.1. Data curation (pre-processing steps)

The applied HTS procedure produced 499,636 raw reads. Pre-processing (data curation) steps were performed with MOTHUR v.1.33.0 (Schloss *et al.* 2009). After merging the forward and reverse reads, we removed all sequences with ambiguous positions (N), those containing homopolymers longer than 30 nucleotides (nt), and any sequence longer than 400 or shorter than 200 basepairs (bp). Chimeric (artificial or naturally occurring recombinant) sequences were detected with UCHIME algorithm (Edgar *et al.* 2011) in *de-novo* mode as implemented in MOTHUR. This method is optimized for detecting chimeric sequences in short, noisy sequences and produces a chimera-free database via an all-against-all pairwise sequence comparison and by exploiting the relative HTS abundance data. We then removed all sequences with abundance ≤ 3 per sample to get rid of possible contaminations, as in Piredda *et al.* (2020). The final dataset included 145,643 HTS reads. A large part of them were identical; a total of 4,693 unique 5S-IGS variants were identified (listed in *Supplementary file S2*).

In contrast to standard HTS data, 5S-IGS data may include reads with potentially artificial sequence patterns that are not filtered for by standard curation procedures applied by MOTHUR. Some of the here compiled sequences reads show imperfect ends (example provided in **Fig. S1**), which hinder the recognition and clipping of the ID-tag + primer sequence regions and may impede correct auto-alignment. Thus, for data completion, we manually checked and clipped 5' and 3' extensions in the 4,693-tip matrix. Folder *4693Data* in the Online Data Archive (ODA) includes both a block-aligned version of the raw data (after being processed and filtered by MOTHUR) and the cleaned alignment used for all analysis.

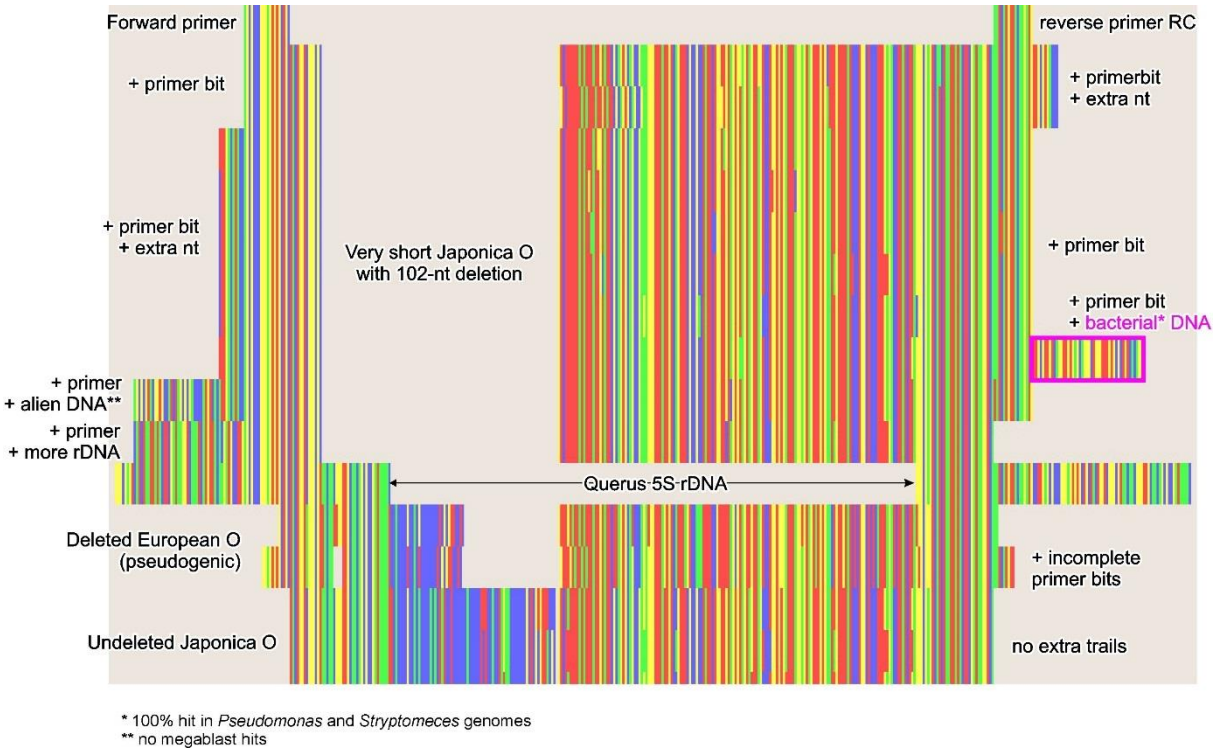


Figure S1 | Bird's eye view of block-aligned data; showing the imperfectly clipped ends in some sequences of short and normal-length type O variants (see **Section 4**). The 5S rRNA genes of *Quercus* ("Quercus 5S rDNA") and the used primer sequences (which usually are clipped automatically) are included for orientation. 'Alien' or bacterial DNA bits flanking the actual amplicons are probably ID-tags, standard pre-processing procedures failed to remove. Curious, but very rare, cases are reads where the primer bit is preceded by genuinely looking rDNA ("more DNA"), i.e. the read includes more rDNA than it should. Such patterns are occasionally encountered in cloned rDNA spacer data as well, and may point to amplification artefacts or pseudogenic, incomplete dimers (see following section).

2.2. Cross-check with 5S rDNA data in gene banks

No beech 5S rDNA or intergenic spacer sequences are currently available gene banks such as the NCBI nucleotide archive (https://www.ncbi.nlm.nih.gov/; last accessed on 25/11/2020). Randomly chosen sequence reads revealed 98–100% identity of the first 40–60 and last 30–50 sequenced bp with the 3' and 5' ends of cloned 5S-IGS sequences of various Fagales: *Quercus* spp. (oaks, Fagaceae; Simeone *et al.* 2018), *Alnus* spp., *Betula* spp., and *Carpinus* spp. (Betulaceae, same order; Forest & Bruneau 2000; Forest *et al.* 2005), as well as with the 5S locus of an ongoing genome sequencing of *Juglans regia* (Juglandaceae, same order; BioProject: PRJNA350852). High (95–100%) identity values were also scored by these two subregions with several other eudicot nuclear-encoded 5S rRNA gene (5S rDNA) sequences, including *Morus alba*, *Punica granatum*, *Vitis vinifera*, etc. The overall mean genetic distances of the entire HTS dataset in these two sequence portions calculated as the number of base differences per site from averaging over all sequence pairs with MEGA X (Kumar *et al.* 2018) were very low (0.01 and 0.02, respectively). We therefore concluded that the two regions belong to the 3' and the 5' portions of the highly conserved 5S rRNA gene; and confirmed this

assessment visually (Supplement file S4; *overview.nex* in ODA). In contrast, no high similarity blast scores were detected for the region comprised between these two portions, the 5S-IGS. This is not unexpected since *Fagus* is genetically most distinct from all other Fagaceae; in coding gene regions the sequence divergence between *Fagus* and the remainder of the family matches situation found between families of core Fagales (**Fig. S2**). The ITS1 of the nuclear-encoded 35S cistron is partly unalignable (Denk *et al.* 2002; Denk, Grimm & Hemleben 2005).

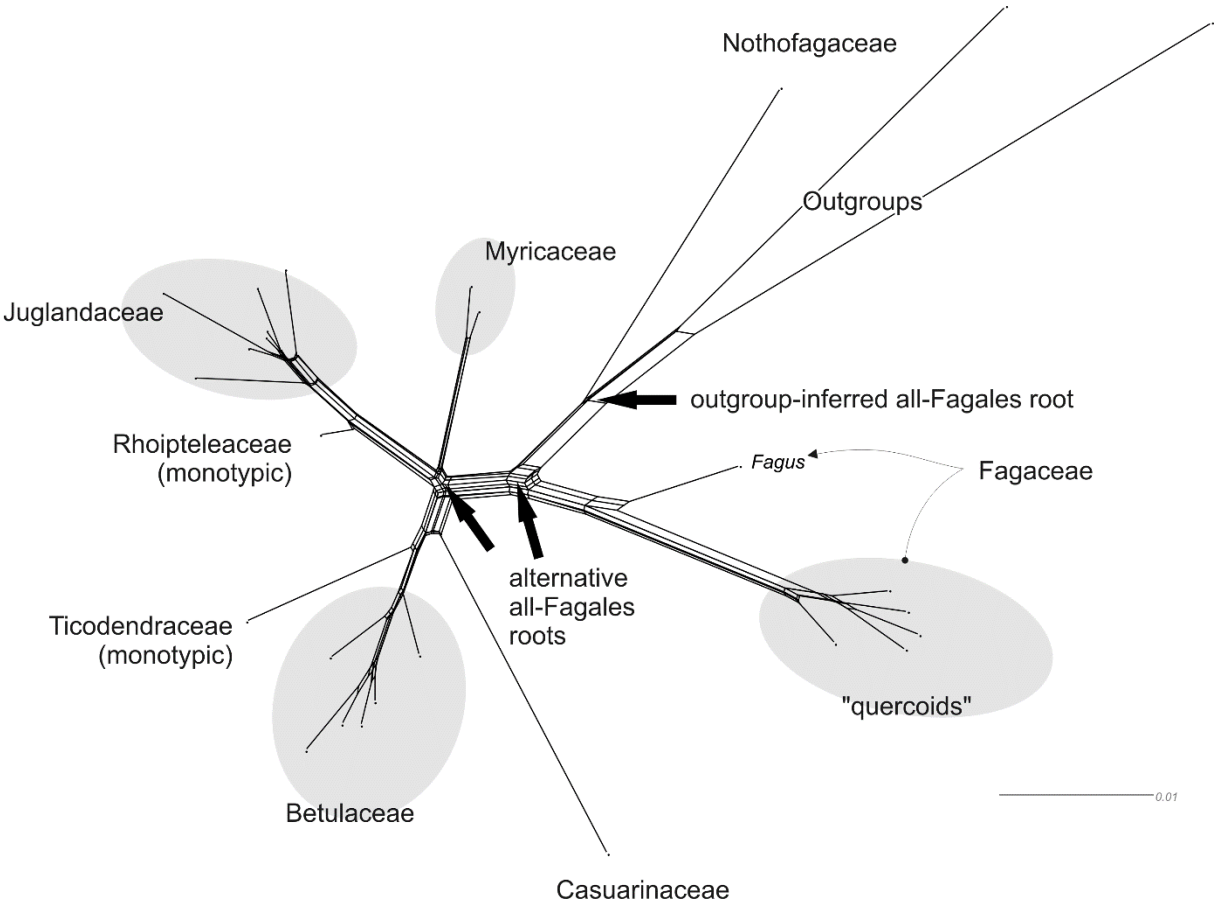


Figure S2 | Neighbour-net for Fagales showing overall genetic divergence within the order (after Grímsson *et al.* 2016, fig. 1). The graph is based on model-optimized genetic distances (scale give number of expected substitutions per site) inferred from the 6-gene matrix of Li *et al.* (2004); for data see Grimm (2020). The genetic difference between *Fagus* and the remainder of the Fagaceae ("quercoids") is comparable with inter-family divergence in the rest of the crown Fagales (note: the current angiosperm classification APG IV includes the monotypic *Rhoiptelea,* Rhoipteleaceae, in the Juglandaceae).

The final, auto-generated multiple sequence alignment (MSA; included in ODA, folder *4693Data*) of the total dataset comprising 4,693 unique sequence variants was 468 bp long. A comparison with published *Quercus* 5S rDNA sequences (e.g. Tynkevich & Volkov 2019) showed the obtained sequences comprise 44 and 34 bp of the upstream and downstream 5S coding region, respectively. Length heterogeneity was due numerous poly-nucleotide motives (mostly poly-T) and indels (insertions, deletions) of variable length occurring in the 5S

intergenic spacer (5S-IGS). The gappyness of the auto-aligned MSA in the flanking upstream gene region is due to a peculiarity found in low-frequent variants that lack a long portion including the end of the upstream 5S rRNA gene and the subsequent 5' part of the intergenic spacer ("short type O"; see below). A few of these "short O" variants show a duplication (genuine or artificial) of the forward primer preceded by a non-5S rDNA four-nucleotide (nt) motif; in one variant the amplicon includes nearly two-thirds of the upstream gene while lacking the 4-nt motif (included in **Fig. S1**). The latter could be an indication that the primer duplication is not an amplification artefact but rather represents an imperfect dimer, where the IGS is eliminated as well as part of the genes.[3]

2.3. Initial labelling and sample-exclusivity of sequence reads

We categorized the obtained 5S-IGS dataset based on their sample distribution (**Fig. S3**; *Supplementary file S2*, sheet 'Representative seq_Reads'). Sequences were labelled as "specific" when exclusively found in the one or two of the samples representing the same taxon: "Japonica", "Crenata", "Iranian orientalis", "Greek orientalis", and "Sylvatica". Four 'Sylvatica' variants, including the overall most abundant one, can be found in Greek *F. orientalis* but with abundances ≤ 2 (<0.001%). Conversely, one 'Greek orientalis' variant was found in *F. sylvatica* from Germany. We identified four additional "ambiguous" classes, i.e. sequences shared among different species or taxa (with an occurrence of 1.3 to 42.7% per sample): 175 variants, classified as "European" and corresponding to 7,271 sequences are exclusively shared between *F. sylvatica* s.str. and Greek *F. orientalis*, while substantially rarer variants (three variants representing 21 sequences) shared across all western Eurasian samples were labelled as "Western". The last two shared sequence classes connect disjunct beech populations: "Ancient" is exclusively shared by *F. sylvatica* s.str. and *F. orientalis* from Iran (three variants representing 45 sequences); "Cross-Asia" is a variant corresponding to 73 *F. crenata* HTS reads and a single sequence of *F. orientalis* from Iran.

[3] Complete (perfect) dimers would have been filtered automatically during the pre-processing being > 400 bp.

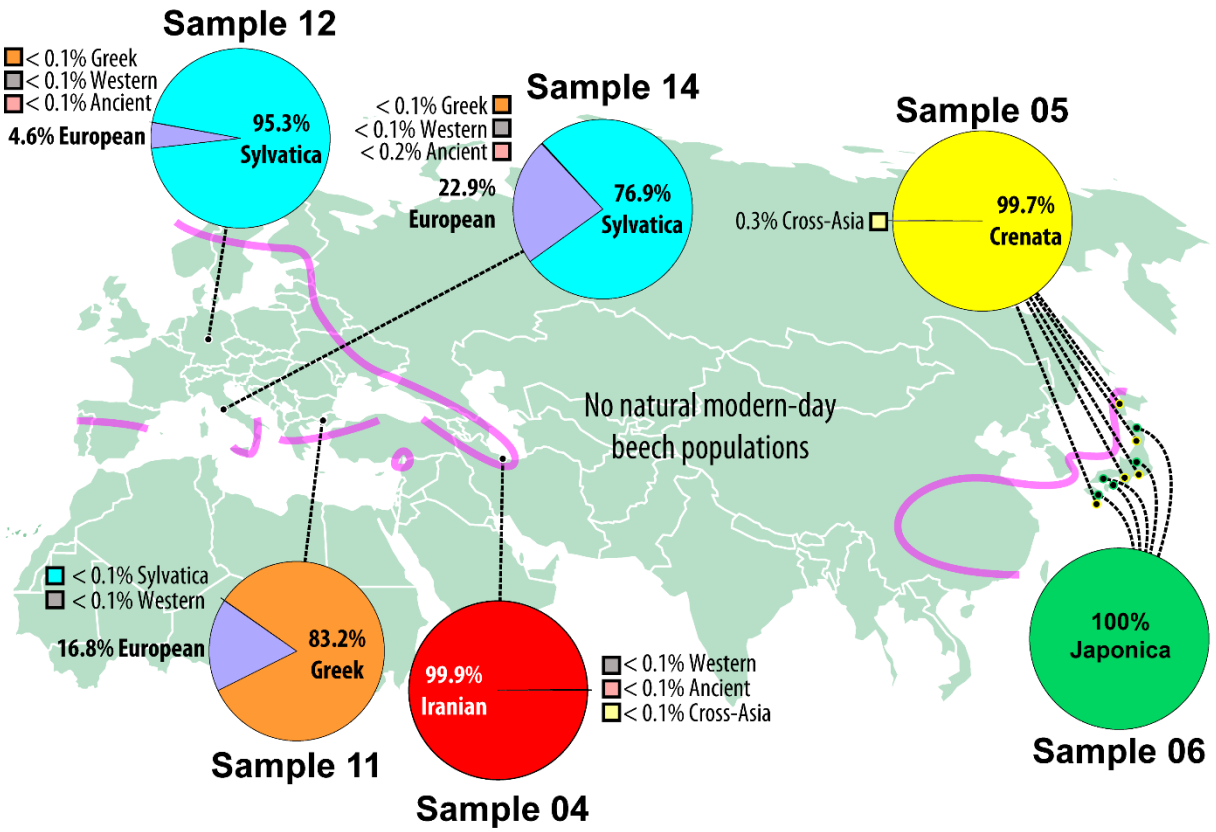


Fig. S3 | Per-sample distribution of "specific" (coloured) and "ambiguous" (grey) 5S-IGS sequence classes. Based on the investigated dataset of 4,693 representative sequences, i.e. non-identical 5S-IGS variants with an abundance of ≥ 4. Purple lines give approximate natural boundaries of beech in western Eurasia and East Asia (after Denk & Grimm 2009, fig. 3); see also https://www.gbif.org/species/2874875.

The *F. japonica* pooled sample included only exclusive sequences (100% 'Japonica'; **Fig. S3**); likewise, *Fagus crenata* and Iranian *F. orientalis* are near-exclusive (>99.5% "specific" sequences). In contrast, *F. orientalis* from Greece has only 83% "specific" sequences, whereas Italian *F. sylvatica* showed the highest percentage of shared ("ambiguous") variants (23%). The highest number of sequences were shared among *F. sylvatica* s.str. (both provenances) and Greek *F. orientalis* (class 'European'); only a few (0.003 – 0.01%) were shared between Iranian *F. orientalis* and *F. sylvatica* s.str., all western Eurasian beeches, or *F. crenata*. The amount of shared and exclusive ("specific") 5S-IGS variants clearly distinguish the geographically isolated Iranian populations of *F. orientalis* from the Greek one, the latter being much more interconnected with nearby *F. sylvatica* s.str.

3. General Features of *Fagus* 5S-IGS

The obtained 5S-IGS sequences were highly variable in structure and length, both at the intra- and interspecific level (**Figs S4, S5**; **Appendices A, B**; *Supplementary file S2*, sheet 'GC content & length'), totalling to an overall mean genetic distance of 0.1. The GC content of the

amplicons ranged from 33.2 to 45.1% (23.9–44.8% for the spacer), and the length range was 166–307 bp (spacer: 88–229 bp). Such a structure and length variability can be found in different plant groups (e.g., Fulnečk *et al.* 2002; Negi *et al.* 2002; Forest *et al.* 2005; Denk & Grimm 2010; Grimm & Denk 2010; Garcia & Kovařik 2013; Mlinarec *et al.* 2016; Garcia *et al.* 2020).

The average GC content is lower than that found in these studies and in the only other genus of Fagaceae studied to date (typical GC content ~50%; Denk & Grimm 2010; Simeone *et al.* 2018; Tynkevich & Volkov 2019; Piredda *et al.* 2020). Interestingly, ongoing full genome sequencing projects of *F. sylvatica* s.str. and *F. crenata* (GenBank acc. no. QCXR00000000.1 and BKZX00000000.1; not yet annotated, accessed on 25/11/2020) reported very low preliminary GC content (34.9% and 32.9%, respectively). *Fagus* further differs from other Fagaceae in general, particularly oaks, by substantially longer ITS regions including more prominent AT-dominated sequence portions and, consequently. lower GC percentages, especially in the ITS1 (Denk *et al.* 2002; Denk, Grimm & Hemleben 2005; Denk & Grimm 2010).

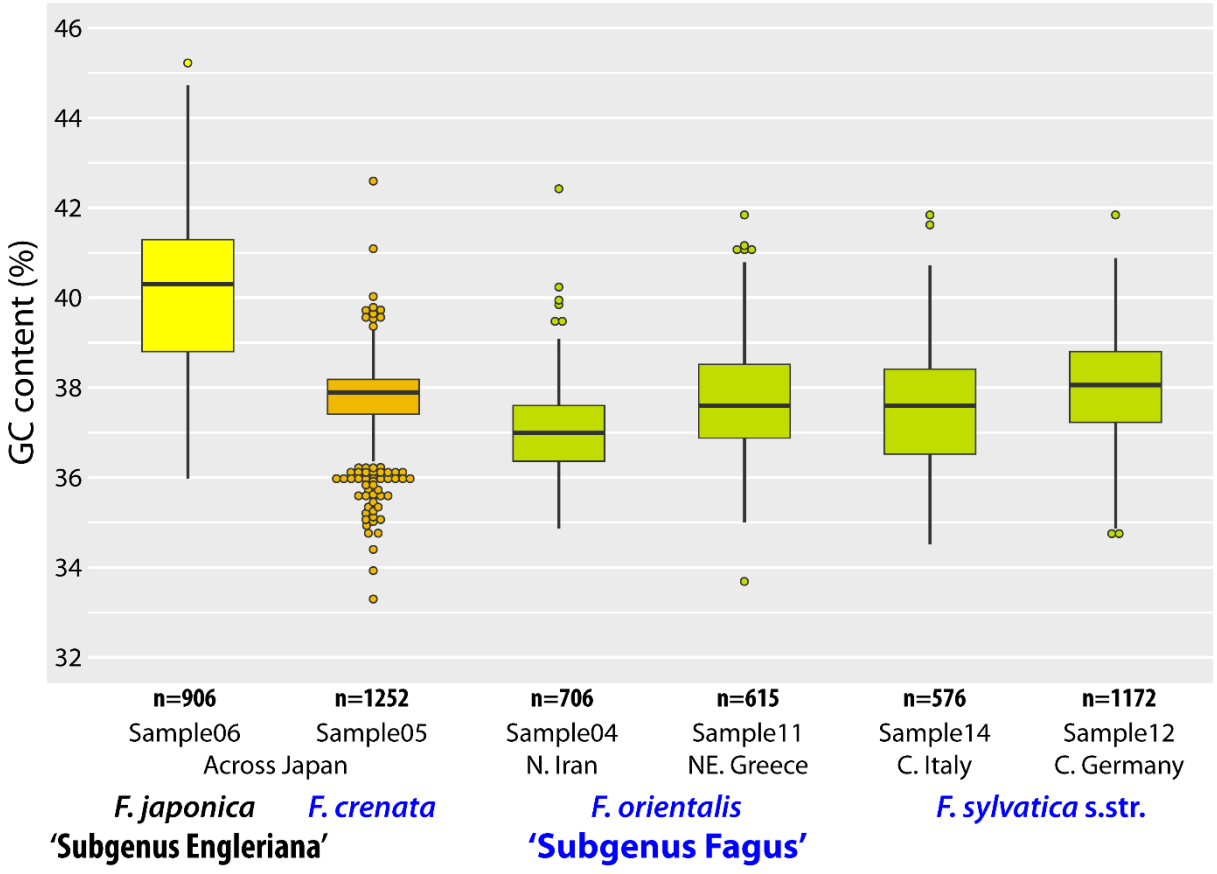


Fig. S4 | Boxplots of amplicon GC contents per sample. Boxes give the median value and the 25th to 75th percentiles, whiskers refer to the entire range without outliers (dots) as defined by Tukey's formula (Tukey 1949).

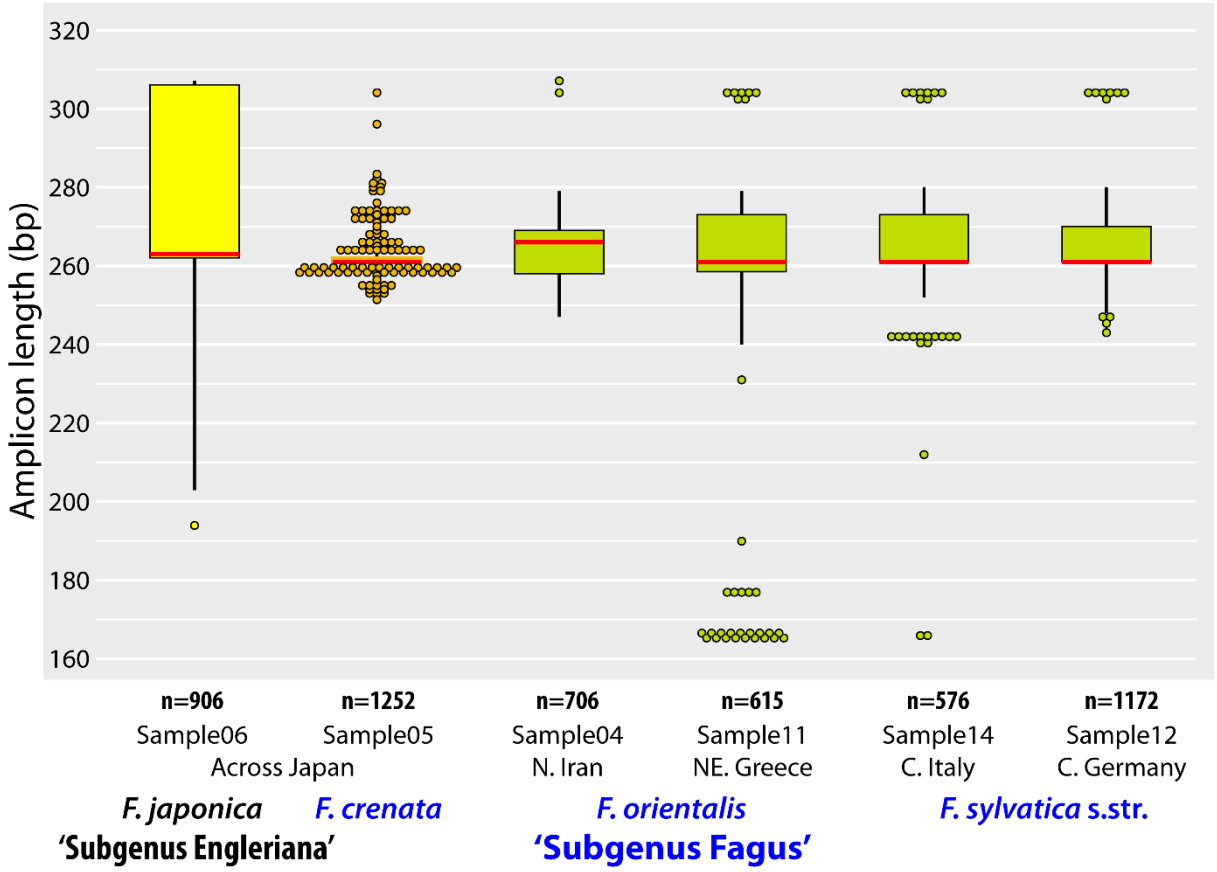


Fig. S5 | Boxplots of amplicon lengths per sample. Boxes give the median value (in red for contrast) and the 25th to 75th percentiles, whiskers refer to the entire range without outliers (dots) as defined by Tukey's formula (cf. **Fig. S4**).

The GC content (**Fig. S4**) and length variation compiled per sample (**Fig. S5**) pointed to a structural difference between 5S-IGS of *F. japonica* ('subgenus Engleriana') and the other taxa (crown group of 'subgenus Fagus'), with numerous sequences showing over-average GC content and length. All other samples were, as a trend, more homogeneous in sequence structure. The GC range for the two combined *F. sylvatica* s.str. samples was 34.4–41.7% (median = 37.9%). In sum, the Iranian population of *F. orientalis* showed the lowest GC contents (median = 36.9%) but highest median length (266 bp).

3.1. Relationship between GC content/amplicon length and abundance

The GC content vs abundance scatterplot of the full dataset (**Fig. S6**) displays a bimodal distribution, with two peaks at 43.6% (*F. japonica*) and 37.9% (*F. crenata*) and abundances of 772 and 7730, respectively. The largest part of the 5S-IGS sequences (98.9% = 144,094 reads; representing 4,616 out of 4,693 non-identical variants) fell within a (>)34–(<)43% range, whereas a second group with higher GC content (>43%) included 74 variants with a total abundance of 1516 reads found exclusively in *F. japonica* (one frequent variant; remaining

individual abundances ranging between 4–60). Three rare variants showed GC content <34% (**Fig. S6**). The length/abundance scatterplot (**Fig. S7**) shows four sequence clusters: (*i*) a relatively low-frequent group of very short sequences (<180 bp); (*ii*) a length-homogenous group (203–204 bp) with low to high abundances; (*iii*) a large cluster of sequences with an abundance maximum around ~260 bp; and (*iv*) a low- to high abundance, length-homogenous group centred around ~305 bp (303–307 bp).

In *F. japonica*, we found all four main length clusters, and a single sequence of 222 bp (with abundance of 16) missing 30 bp of central IGS. The very short (≤ 170) and 204-bp cluster (102 5S-IGS variants represented by a total of 1865 reads; 5.8% of this species'/sample's total) share a 102 nt-long deletion involving 31 bp of the 3' end of the 5S rRNA gene (starting at invariable alignment pos. 14 and represented by "Ja_short" in Supplementary file S4, sheet *Selected variants*); all but one belong to the group showing a GC content ≥ 44%.[4] The most length-homogeneous sample was *F. crenata* (**Fig. S5; Table S1**).

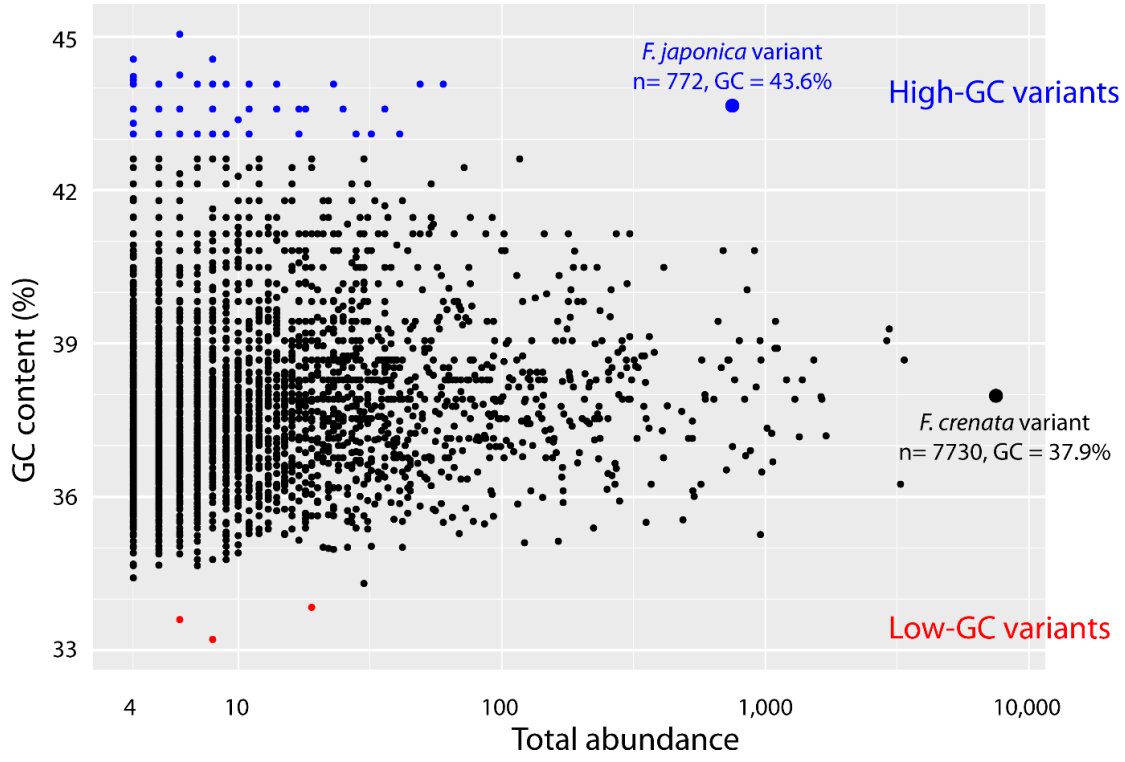


Fig. S6 | x-y plot of the GC content and abundance of the 4,693 5S-IGS sequences. Shown is total abundance, i.e. the sum across all samples (cf. *Supplementary file S2,* sheet 'GC content & length').

[4] The sequence with lower GC content belongs to a different main 5S-IGS lineage (Japonica type I; see ***Phylogenetic sorting***)

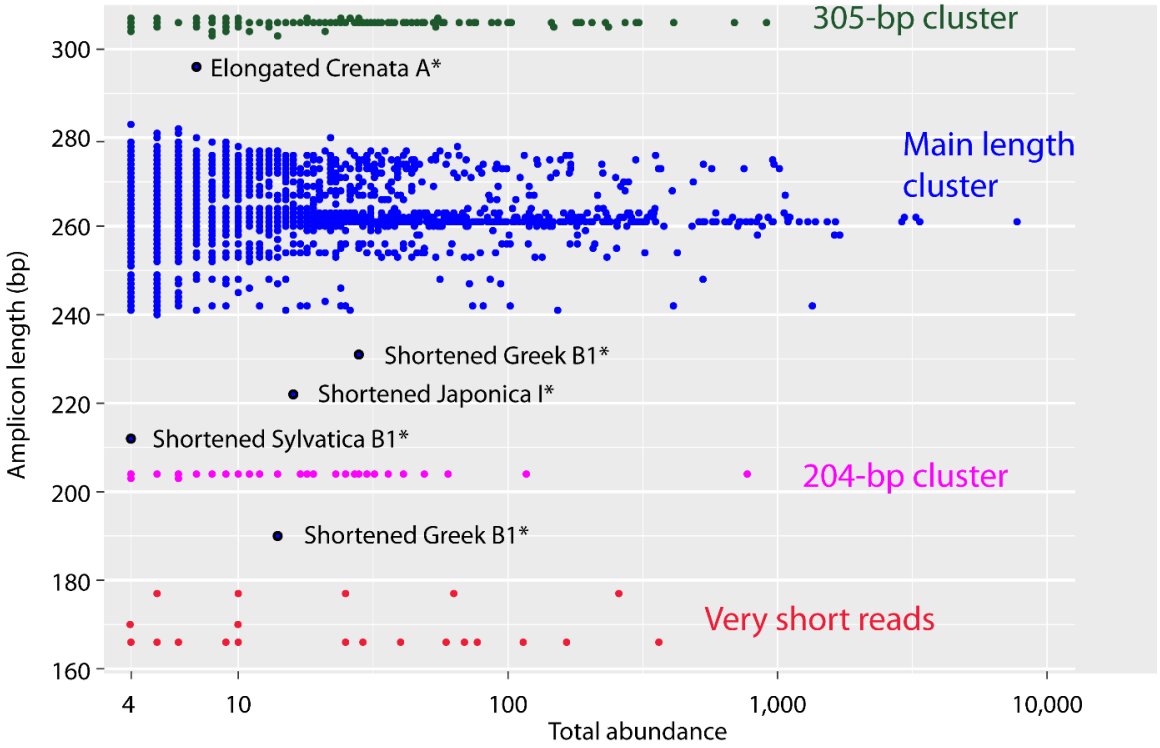


Fig. S7 | x-y plot of the length and total abundance. Downstream phylogenetic analysis (cf. main-text figs 3–5; Supplement file S2) shows that sequences with a length of ~280–300 bp either represent shortened versions of the 305-bp cluster or elongated versions of the main length cluster. * For labels see section ***Phylogenetic Sorting*** below, and main text.

The Iranian *F. orientalis* population showed a higher diversity in sequence length and GC contents (**Figs S4, S5)**. Most sequences of *F. sylvatica* s.str. from Italy (99.6%; 565 variants) and *F. orientalis* from Greece (92.4%; 583 variants; **Table S1**), the samples with the most pronounced length polymorphism (**Fig. S5**), fell within the main length cluster. Most of the observed length variation within the main length cluster relates to the co-occurrence of two main 5S-IGS lineages in samples representing *F. sylvatica* s.l. (cf. main-text fig. 7) of ~240–270 (type B lineage) and ~260–280 bp (type A lineage; see below). Very few sequences of *F. sylvatica* s.l. (in total 10 variants with a sum-abundance of 100, i.e. ~1 ‰ of all reads) have a length ≥ 303 bp. The group of very short variants—23 "Greek-orientalis specific", one "Sylvatica-specific", one "ambiguous" variant, 166 or 177 bp long with abundances of 4–363—show an identical ~95 nt-long deletion (represented by "OrG_short" in Supplement file S4, sheet *Selected variants*). One variant belonging to the Greek sample of *F. orientalis* (length = 231 bp; GC content = 36.4%; abundance = 28) had a 30 nt-long deletion involving 29 bp of the 3' end of the (upstream) 5S rRNA gene. As in the case of the short *F. japonica* variants (204-bp length cluster), the very short *crenata-sylvatica* lineage variants and the variant missing the end of the upstream 5S rRNA gene show no further evidence of sequence degradation. The

downstream 5S rDNA bits are inconspicuous in showing the genus'/family's consensus sequence.

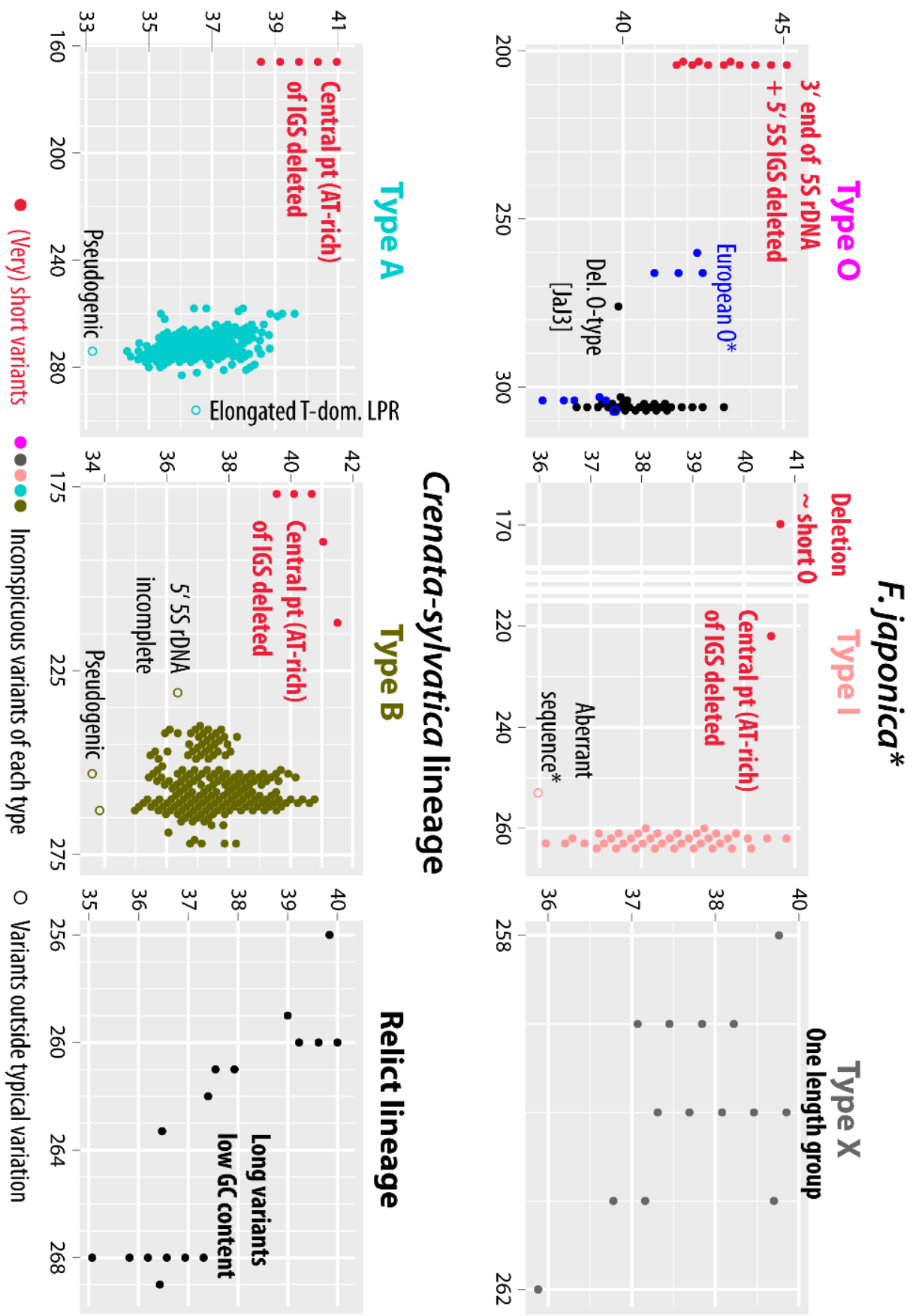
Table S1. Number of non-identical sequence variants (NIV) and total abundance (TA) of length clusters (as defined in **Fig. S7**).

	F. japonica		*F. crenata*		Iranian *F. orientalis*		Greek *F. orientalis*		*F. sylvatica* s.str.	
Length cluster	NIV	TA	NIV	TA	NIV	TA	NIV	TA	NIV	TA
Very short	3	18	–		–		23	1345	2	12
203–204 bp	99	1847	–		–		–			
Main length cluster	410	20785	1250	27364	704	20213	583	17100	1227	18303
…range	258–276 bp		252–283 bp		247–279 bp		240–279 bp		241–280 bp	
~305 bp	393	9559	1	4	2	14	7	20	9	66
Total	906	32225	1252	27375	706	20227	615	18507	1239	18385

4. Phylogenetic Sorting

The subsequent sorting of the obtained dataset into five main phylogenetic lineages (*F. japonica* types O, I, and X; *F. crenata – F. sylvatica* s.l. types A and B; cf. main text, *Supplement files S2–S4*) revealed a structural homogeneity of the sequences included in each type (**Fig. S8**; **Appendix A**; see also main-text fig. 7). The *crenata-sylvatica* lineage A-types have generally lower GC content and are shorter than the B-types in each species/sample; the I-and X-types (*F. japonica*) correspond regarding CG content and length ranges to type B of the other taxa. Type O (*F. japonica*) includes longest 5S-IGS variants (typically ≥ 300 bp; 305 bp-class) and highest GC contents.

(Following page) Fig. S8 | x-y plots of GC content vs amplicon length; sorted for the five main phylogenetic types (5S-IGS lineages) and 5S-IGS variants forming the 'Relict lineage'. Dots may represent more than one variant, since variants may have (near-)identical GC contents and lengths. "Aberrant sequence" type I sequence ("SyG3" in *Supplementary file S4*) is a sequentially unique, weakly degraded variant obtained from the German *F. sylvatica* sample without affinity to any major type and placed within type I subtree by EPA (Evolutionary Placement Algorithm). * Low-frequent (total abundances ≤ 36) O-type variants (addressed as 'European O'; see **Section 4.3**) can be found in all samples of the *crenata-sylvatica* lineage.


F. japonica*
Type O
3′ end of 5S rDNA
+ 5′ 5S IGS deleted
European O*
Del. O-type [JaJ3]
Deletion
Type I
~ short O
Central pt (AT-rich) of IGS deleted
Aberrant sequence*
Type X
One length group
Crenata-sylvatica lineage
Type A
Central pt (AT-rich) of IGS deleted
Pseudogenic
Elongated T-dom. LPR
Type B
Central pt (AT-rich) of IGS deleted
5′ 5S rDNA incomplete
Pseudogenic
Relict lineage
Long variants low GC content
(Very) short variants
Inconspicuous variants of each type
Variants outside typical variation

4.1. Sequence features in I-, O- and X-types of F. japonica

In *F. japonica*, the three distinct and conserved length classes (**Table S1**) sort phylogenetically. I- and X-types have similar length (~260 bp; main length cluster) and GC contents; O-types have higher GC contents and form the 204- and 305-bp length clusters (**Fig. S9**). The short (203–204 bp) sequences represent truncated type O sequences ("Short O" in *Supplement file S4*), missing 30 bp of the upstream 5S rRNA gene and the 5' end of the 5S-IGS up till and including all but one nucleotide of the T-dominated 5' length-polymorphic region (LPR; *Supplement file S4*, sheet 'Motives'). While such a deletion is characteristic for pseudogenic, dysfunctional rRNA arrays/copies, the remainder of the sequences including the downstream 5S rDNA are inconspicuous, hence, the very short type O sequences show the highest observed GC contents (**Fig. S8**).

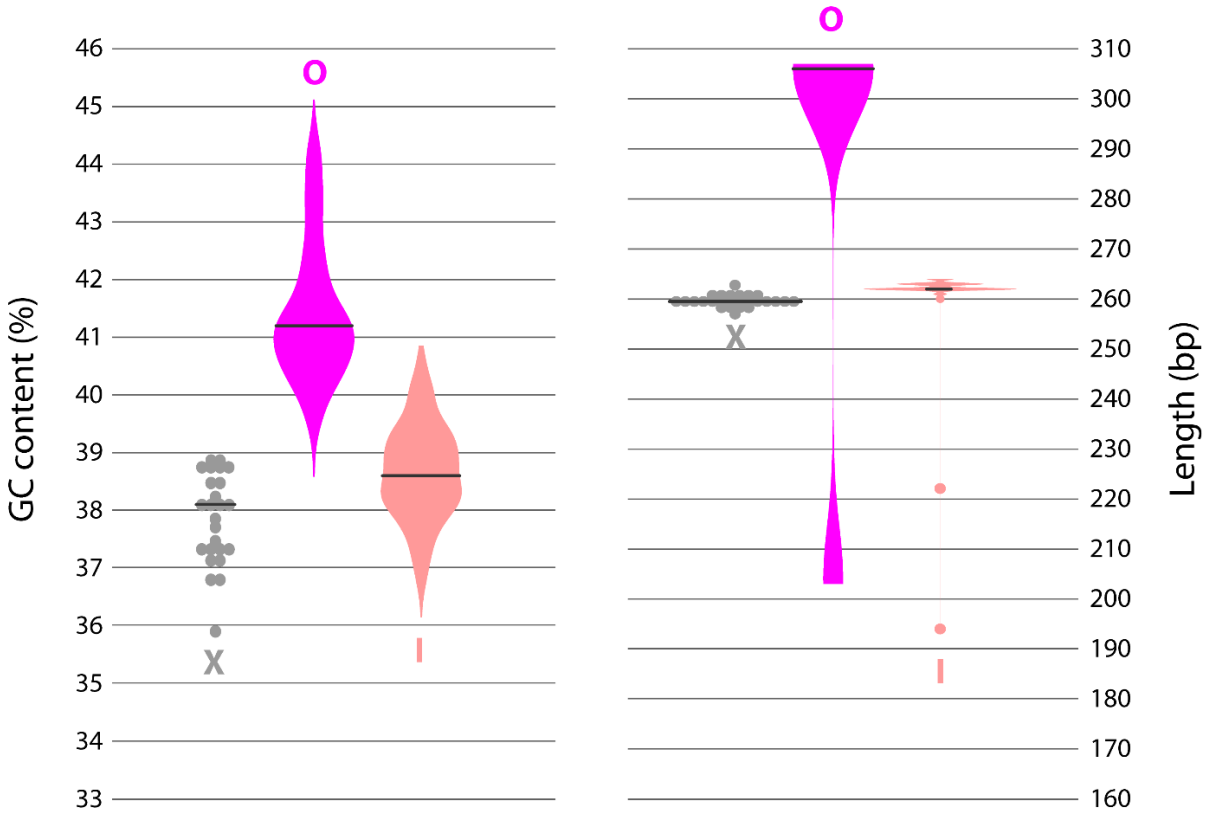


Fig. S9 | Violin and scatter plots of GC content and sequence length for three 5S-IGS lineages observed in *F. japonica*. Type O represents an outgroup 5S-IGS type (see main-text), type X a non-distinct ingroup type, and type I the sisterlineage of *crenata-sylvatica* type B.

The most apparent diagnostic sequence feature of type O is a much-elongated central length-polymorphic region, the "semi-homologous region"[5] defined in *Supplementary file S4*, and a generally high GC content in the length-homogenous portions of the spacer. The sequential difference to the 'ingroup' A-, B- and I-types is striking, allowing to quickly identify type O variants when viewing (auto-generated) alignments in bird's view[6]. The 3' variable region following downstream of the semi-homologous region shows an even increased amount of point mutations, mostly transitions leading to a higher GC content than in A-, B-, I- and X-types.

In contrast to the O and I types, the rare type X seem to lack discriminative sequence patterns obscuring its relationship with respect to the A-type and B+I-type lineage (exemplified in **Fig. S10** for the 5' T-dominated LPR). The 686-tip tree placed X-type variants as first diverging lineage within the 'ingroup' clade; its ambiguous, non-discriminate signal caused the collapse of branch support for deep (inter-type) relationships (main-text fig. 3). Sequentially, it clearly represents an 'ingroup' sequence, which however lacks any clear diagnostic or phylogenetically informative feature in contrast to A- and B-/I-types. When deviating from the ingroup consensus, i.e. the modal consensus of type A, B, I and X variants, it either shows sequence patterns diagnostic for the A-type lineage, the B+I clade or found as intra-lineage variation in either of them (**Table S2**). The lack of decisive phylogenetic sequence patterns is the main reason for the low support in the 486-tip tree (main-text fig. 3); thus, we did not include the type for the select 36-sequence matrix (main-text fig. 5). Without more data on East Asian beech species, the (sequence-wise) rather primitive, underived *F. japonica* type X must remain an enigma. Possible evolutionary scenarios for type X include:

- Ancient variant, originally shared by all lesser-evolved East Asian species, i.e. left-over of incomplete lineage sorting.
- A common variant in not sampled Chinese beech species, i.e. evidence for (past) reticulation.
- A rare relict variant exclusive of 'subgenus Engleriana', largely replaced by I-type variants in *F. japonica*.

[5] While position homology is straightforward to establish within each main-lineage, alternative alignments are possible between main lineages within this generally length-polymorphic and sequentially high-divergent part of the 5S-IGS. While the semi-homologous region (SHR) as a whole is homologous between "outgroup" type O and "ingroup" types, its is unclear which bit of the much-elongated SHR motif of type O corresponds to bits seen in "ingroup" types.

[6] The ODA includes a NEXUS version of the auto-aligned total 5S-IGS data, annotated with and optimised for viewing with MESQUITE, which provides a bird's eye view option (Maddison, W.P. & Maddison, D.R. (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75.)

Table S2 | Phylogenetic ambiguity of type X sequences. Matches are highlighted by bold font and colour. IUPAC polymorphism codes used when there are co-dominant motives within each lineage (i.e., reflect plurality consensi; for complete documentation see *Supplementary file S4*)

Type	Pos. 81ff	T-dom. LPR (96–218)	135ff	147f	157	SHR[a] (174–241)	275
Affinity of X	→ B	→ B	Outgroup	None	Ingroup[b]	Ingroup	Ambiguous[c]
X	TT**T**-TATA	**"Ancient B"**	**CAA**[d]	TT	**C**	**"Ancestral"**	**T**
A	TTG-TATA	"Long"	CTT	RA	**C**	**"Ancestral"**	A
						"Asian types"	
I	TTG-TATA	"Ja intype"	CGA	GA	T	**"Ancestral"**	**T**
B	TT**K**-TATA	**"Ancient B"**	CGA	RA	**Y**	**"Ancestral"**	**T**
		"Origianl B"	YTA			"Asian types"	
		"European B"	TCG			Various derived types	

[a] Semi-homologous region

[b] Usually candidate for an ancestral sequence feature within the ingroup; subsequently lost or modified in the ingroup sibling lineages.

[c] Highly conserved SNP: always T in 'Japonica O', types B and I; consistently A in type A and 'relict lineage' variants.

[c] Similar to Type O: AAA, **CAA** in 'European O'

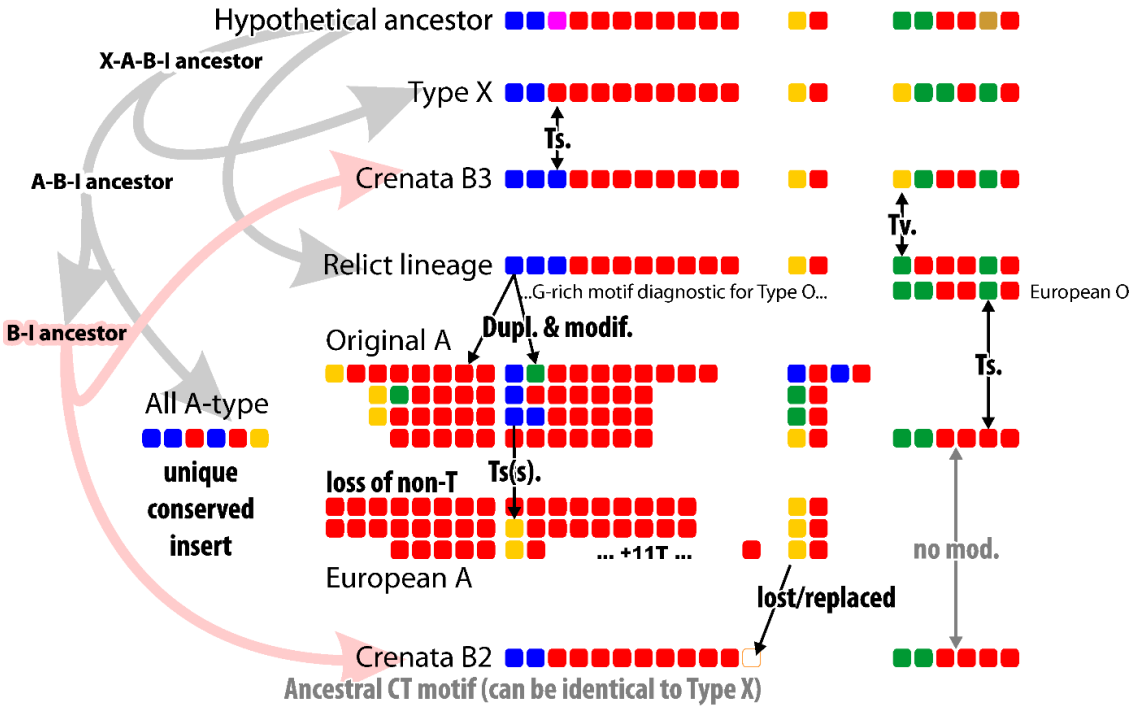


Fig. S10 | Hypothetical (early) evolution of the 5′ T-dominated length-polymorphic region in the 5S rDNA intergenic spacer of beeches, leading to low-modified B-subtypes and A-type diagnostic elongation (see also *Supplementary file S4,* sheet 'Motives'). Coloured squares represent nucleotides: green = adenine (A), blue = cytosine (C), dark yellow = guanine (G), red = thymine (T); mixed colours uncertainties (hypothetical ancestral sequence) or polymorphisms. Empty squares reflect (here: 1-nt) length polymorphism. The phylogenetic relationships between the shown types follow the result of the 686-tip (main-text figs 3, 4) and 38-tip dataset analyses (main-text fig. 5). Abbrev.: dupl. = duplication; Ts. = transition; Tv. = transversion.

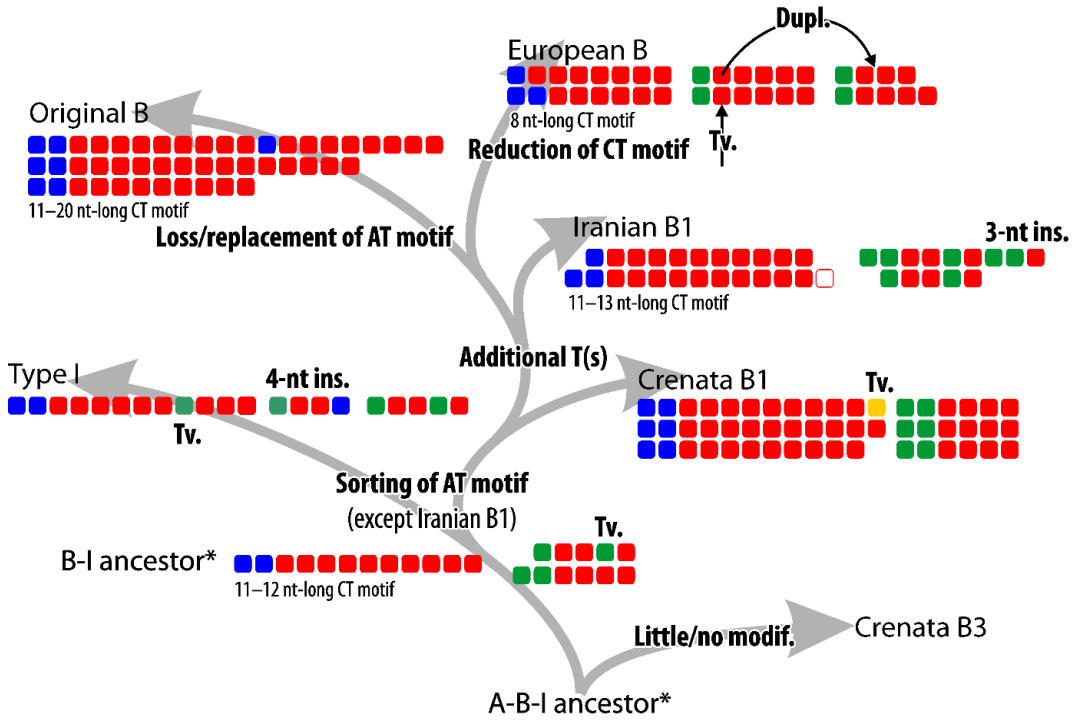


Fig. S11 | Differentiation of the 5′ T-dominated length-polymorphic region (5′ LPR) within the I-B lineage. Same colour coding and abbreviations as in **Fig. S10.** * Crenata B1 variants, accounting for > 80% of *F. crenata* 5S-IGS reads, may show a potentially ancestral 5′ LPR oligonucleotide motif, while the 5′ LPR of very rare Crenata B3 variants lacks the otherwise shared, B-I-lineage-diagnostic features (cf. **Fig. S10**).

The I-type, the most dominant of the *F. japonica* types (**Appendix A**), matches in GC content and length diversity the *crenata-sylvatica* type B (**Appendix B**). Sequentially, it represents the *F. japonica* counterpart to the *crenata-sylvatica* type B as well, sharing several highly diagnostic sequence features differentiating between A-type and B-type variants (*Supplement file S4*). For instance, its T-dominated 5' LPR is sequentially most similar to that of putatively derived B-types such as 'European B' and 'Iranian B1' (**Fig. S11**). It differs from both types by several conserved point mutations, mostly transversions, in the lineage-discriminating central part of the 5S-IGS, in-between the T-dominated 5' LPR and the semi-homologous region.

4.2. Sequence features in F. crenata – F. sylvatica s.l. lineage (A- and B-types)

Downstream analysis of 38 select variants showed that much of the observed length variation relates to the co-occurrence of two main 5S-IGS lineages, type A (~265–275 bp) and type B (~240–260 bp; **Fig. S12**). The split between the two lineages is exemplarily reflected in the motives of the T-dominated LPR at the 5' end of the 5S-IGS (**Fig. S11)**. Conserved point mutations characteristic for the A-type lineage are concentrated directly downstream of the T-dominated 5' LPR. One Crenata A sequence falls out of the usual length range (marked in **Fig. S8**) because of a much-elongated T-dominated LPR showing multiple repetitions of upstream T_x-C and downstream A-T_y oligonucleotides (x = 2–3; y = 4–7) connected by five T. Sequences with ~300 bp belong to the same lineage as the *F. japonica* O-type ('European O'; see **Section 4.3**).

In addition, sample #11 (Greek *F. orientalis*) includes 18 variants ("OrG_short" in *Supplement file S4*) with a similar-located 107-nt (type A)/ 86-nt long (B) deletion in both A- and B-type lineages starting with the T-dominated 5' LPR and stopping before (B-types) or within (A-types) the semi-homologous region. As in the case of the short O-type sequences, pseudogenic mutations are extremely rare in the flanking rDNAs. Since these short variants ('very short' length cluster; **Section 3.1, Fig. S7**) deleted most of the lineage-discriminating central part of the 5S-IGS, they are ± equally distant to both A- and B-type variants, hence, collected in the central part of the neighbour-net splits graph (main-text fig. 4). Their assignation to A- and B-type lineages using EPA or detailed sequence-inspection is straightforward because of the lack (A-types) or occurrence (B-types) of lineage-specific, subtree-sorted point mutations in the non-deleted, length-homogenous but lineage-divergent 3' variable region and the undeleted part of the semi-homologous region (*Supplementary file S4*, sheet 'Motives').

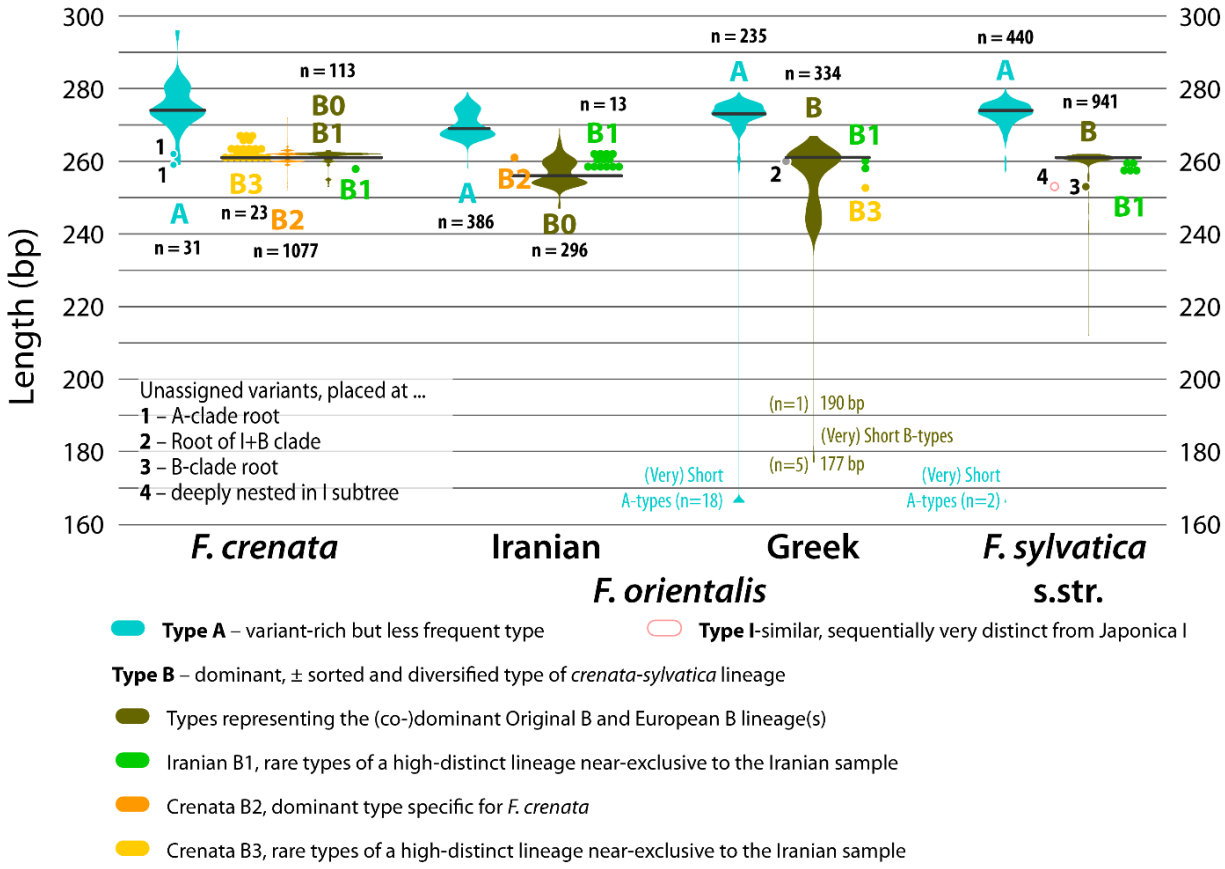


Fig. S12 | Violin and scatter plots of amplicon lengths; shown are 5S-IGS types characteristic of the *crenata-sylvatica* lineage (**Section 4.2**).

B-type variants show a higher GC content than A-type variants (**Fig. S13**) and a much higher oligonucleotide motif diversity in the two generally length-polymorphic regions, the T-dominated 5' LPR (**Figs S10, S11**) as well as the semi-homologous region. Both length-polymorphic regions can be highly diagnostic for dominant subtypes such as 'European B' and 'Crenata B2'. In-depth analysis of the select set as well as visual inspection of the 4,693-sequence alignment (NEXUS-file included in the ODA, folder *4963Data*) indicates that A-type variants include not only longer (always AT-rich) length polymorphic regions but also more variants with an increased proportion of point mutations that could be related to (beginning) pseudogeny, hence, the trend towards lower GC contents. This agrees with the overall abundance pattern of a very few, typically low-GC A-type variants in *F. crenata* and increasing dominance of type B variants from the Iranian to the German population. Overall, the pattern points to an ongoing elimination (silencing) and replacement of (more primitive) A-type by (more derived) B-type arrays in the genomes of the *crenata-sylvatica* lineage (discussed in main text).

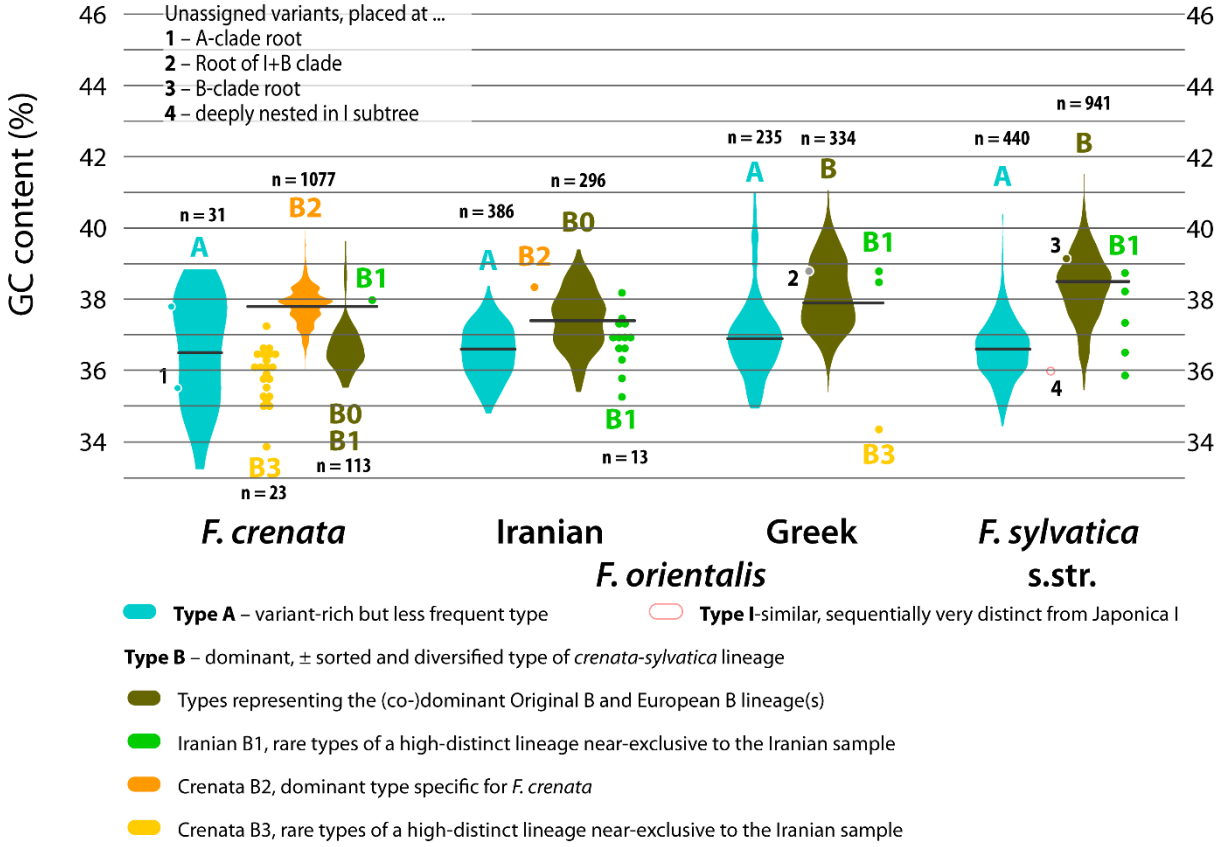


Fig. S13 | Violin and scatter plots of GC contents; shown are 5S-IGS types characteristic of the *crenata-sylvatica* lineage.

4.3. Relict types and taxonomic mismatches

In addition to the main (co-dominant), phylogenetically and taxonomically ± sorted 5S-IGS types, rare variants can be found in all members of the *crenata-sylvatica* lineage that show a stronger affinity to the variants of the outgroup sample and corresponding amplicon lengths and GC contents (*F. japonica*; **Fig. S14**).

***European O lineage*—**Variants labelled as "European O" show all distinctive characteristics of the Japonica O type such as the much-elongated, GC-rich semi-homologous region (SHR). The SHR of 'Japonica O' and 'European O' variants are sequentially not identical but structurally corresponding (*Supplementary file S4*); highest number of distinguishing point mutations are concentrated in a 24 nt-long stretch at the 3' end of the SHR. All European O type variants show an increased amount of putatively pseudogenous transitions from G→A and C→T, in both the spacer and the flanking gene regions. Only a single 'European O' variant is relatively frequent (total abundance of 36; shared by both *F. sylvatica* s.str. samples; "D_ASOG" in the 38-tip dataset), hence, included in the 686-tip dataset, and resolved as sister

to 'Japonica O' clade (cf. main-text figs 3, 5; assuming a midpoint root). 'European O' type variants show a minor (~40 nt-long) length polymorphism: in the shortened variants (such as D_ASOG) the part directly upstream of the (shortened) T-dominated 5' LPR is deleted. 'European O' type variants are more frequent and diverse in the European samples of the *crenata-sylvatica* lineage (**Fig. S14**), with only three variants found in Iranian *F. orientalis* (total abundance, TA, of 24; 0.12% of all post-processing reads; two of class "specific", one "ambiguous" shared with European samples) and *F. crenata* (TA = 30; 0.11%; all "specific").

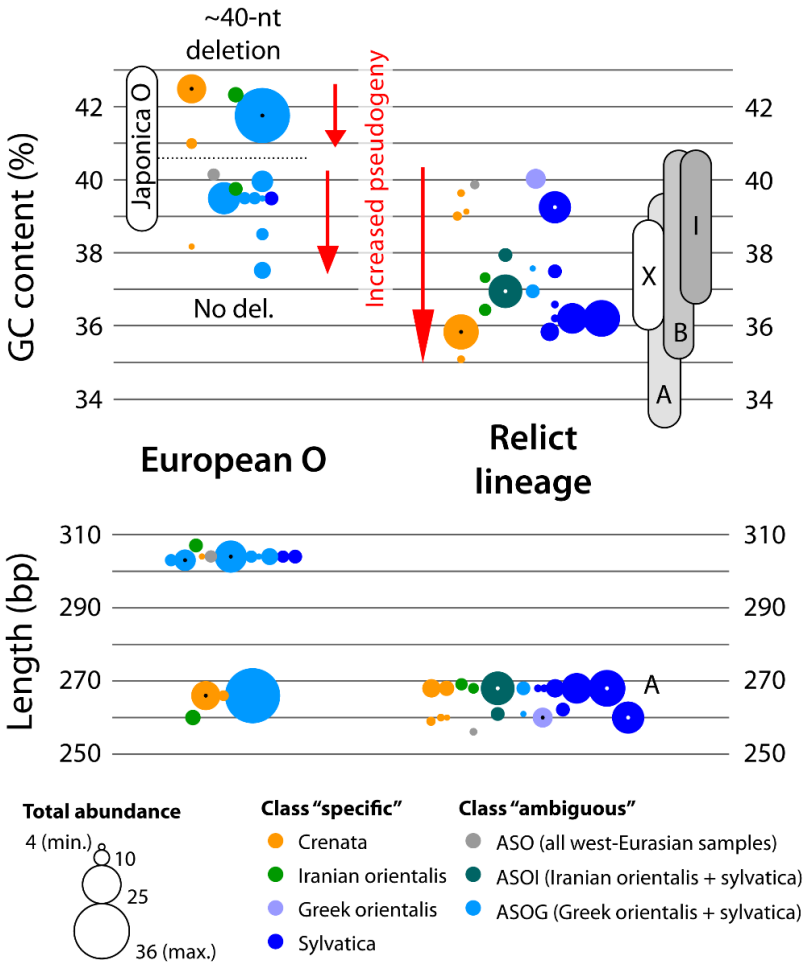


Fig. S14 | Scatter plots of GC content and sequence length of European O and relict lineage variants. Variants classified as "specific" are taxon-restricted, "ambiguous" are shared among studied samples. Bars give the GC range of normal-length variants (main length cluster, 305-bp length cluster in **Fig. S6**) for the major 5S-IGS types in *F. japonica* (types I, O, and X) and the *crenata-sylvatica* lineage (types A and B).

***Ancestral polymorphism*—**The occurrence of a (largely?) degraded sister lineage of the 'Japonica O' type in the *crenata-sylvatica* lineage in addition to the I-B sister relationship is evidence for an ancient 5S rDNA polymorphism already present in the common ancestor(s) of all (Eurasian) beeches, i.e. the last common ancestor of 'subgenus Fagus' in Eurasia and

'subgenus Engleriana' (main-text fig. 8). 'European O' type arrays are probably much more common than covered in our data but not amplified because of high-pseudogenic gene regions with corrupted primer-attachment sequences; while their non-degraded counterparts in the *F. japonica* genome are much easier to capture using the here applied HTS approach (**Section 5.1**). According to fossil evidence and on the background of molecular phylogenies, this polymorphic ancestor can be considered to have lived at least ~35 myrs (cf. Grímsson *et al.* 2016) up to >50 myrs ago (Renner *et al.* 2016); *Fagus* 5S rDNA arrays have hence the capacity to retain independently evolving, alternative tribes of copies for tenth of millions of years.

The sole German-sample variant placed by EPA within the I-subtree (deeply nested; **Fig. S15**) is sequentially very distinct and lacks most (except for one point mutation) sequence patterns diagnostic for the I-lineage ("SyG3" in *Supplementary file S4*).[7] It includes sequence patterns otherwise exclusive to the A- or B- lineages, as well as shared, putatively underived oligonucleotide motives of the A and I lineages. Subtracting point mutations due to beginning pseudogeny, the variant may well represent a leftover of the early diversification into I-, A- and B-types or even an ancient recombinant.

Ancestral polymorphism coupled with incomplete lineage sorting (and secondary mixing) is the only scenario to explain genetic relicts such as 'European O' type and the single putative I-lineage variant found in the German *F. sylvatica* s.str. sample. Sample contamination as alternative explanation can be ruled out insofar that the material was collected and DNA extracted at different times and the lack of identical variants in *F. japonica*, *F. crenata* and Iranian/European samples. *Fagus japonica* is already rare in its native area, and extremely difficult to find in European arboreta or parks, and our west-Eurasian samples come all from natural (and old) beech stands. Hence, introgression because of cultivation can be ruled out as well. Notably, EPA did not recover any 'European O' variant in *F. japonica*, or 'Japonica O', or 'Japonica I', types in any other sample.

[7] This 'not-I' variant exemplifies the risk of mis-assignations when using automatic identification approaches such as EPA without post-analysis cross-check of surprising hits and visual inspection of results. The relatively low likelihood weight ratio for the best-placement alone is not conspicuous given the structure of our data and guide tree (many stochastically distributed mutations, flat terminal subtrees with high number of leaves); but in the case of the 'not-I' *sylvatica* variant, this it is coupled with a very long terminal branch in the EPA placement tree (all EPA placement trees are including in the *Online Data Archive*).

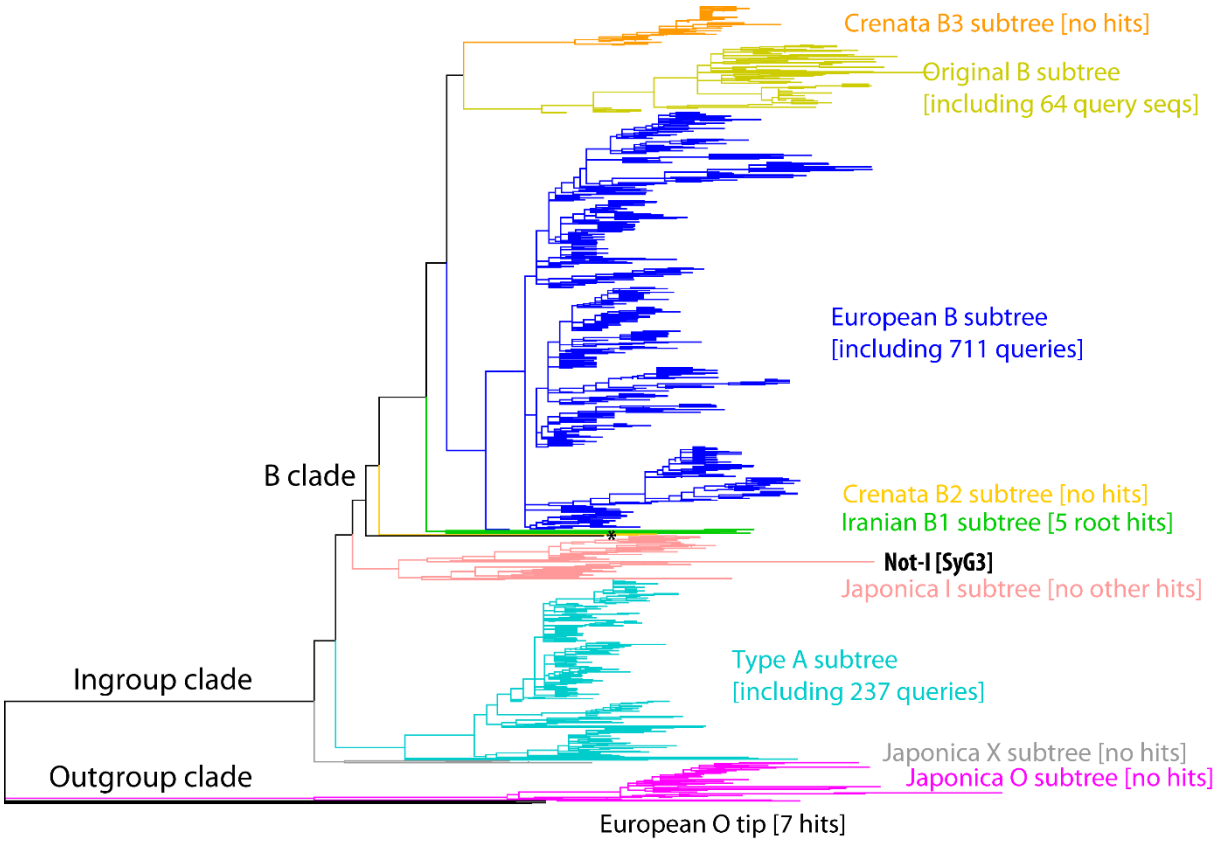


Fig. S15 | EPA placement tree for queries classified as 'Specific sylvatica'. Shown is a phylogram, which gives the topology and branch lengths of the used reference tree (686-tip tree, cf. main-text fig. 3), and the tip-lengths of placed queries (cf. jPlace file included in Online Data Archive). The distinctness of the single query placed within the type I subtree is obvious from such a graph; queries with much elongated tip branches may represent pseudogenic sequences, recombinants, or misplacements of unique sequence types not covered in the reference tree. * One query was placed at the root of the B-clade; with respect to the data and phylogenetic structure of the reference tree, such a placement can be indicative for an ancestral variant that lacks lineage-sorted and diagnostic sequence patterns.[8]

The Relict lineage—In addition, we captured rare variants that are sequentially intermediate between the 'outgroup' Type O and the 'ingroup' types (X, A, B, and I) in all samples of the *crenata-sylvatica* lineage (but not *F. japonica*). Two of these variants are shared by disjunct Iranian *F. orientalis* and *F. sylvatica* s.str. and samples (ASOI class in **Fig. S14**), one found in all west-Eurasian samples. Even more then the European O variants, they can be enriched in potentially pseudogenous point mutations, in both the flanking gene regions and the non-coding, non-transcribed intergenic spacer. With respect to their phylogenetic position between 'outgroup' and 'ingroup', their pseudogenic tendency, and the oligonucleotide motives found in the 5' T-dominated LPR and the semi-homologous region, we refer to them as "Relict lineage". The pseudogenous mutations inflict long-branches in phylogenetic trees (long edges in networks). However, these stochastic patterns did not eliminate and obscure sequence

[8] When using EPA on short reads or queries with missing data, placements at roots can also indicate the query simply lacks the informative, lineage-discriminating sequence portions.

patterns that appear to be conserved within and exclusive to this 5S-IGS 'shadow' lineage (main-text fig. 5; *Supplementary file S4*). Hence, the high capacity of EPA to correctly identify members of this lineage based on only a single potential target in the 686-tip tree. Our hypothesis is that the 'Relict lineage' represents leftover imprints from early divergences or reticulations that much predate the formation of modern species, or involve species not sampled so far (continental and insular East Asian spp.; North American *F. grandifolia*).

5. Comparison with Other Fagaceae/Fagales

Our data revealed five main phylogenetic lineages within the 5S-IGS gene pool of *Fagus*: types O, I, and X in *F. japonica*; types A and B in the *crenata-sylvatica* lineage. In addition, we recovered copies from a likely pseudogenic lineage, the 'Relict lineage'. Intra-specific to intra-genomic 5S-IGS sequence polymorphism (5S-IGS "paralogy") is a common and long-known feature of Fagales (*Corylus*, Betulaceae: Forest and Bruneau 2000; *Quercus petraea, Q. robur*, Fagaceae: Muir et al. 2001). Cloned sequence data demonstrating the extent of 5S-IGS polymorphism are further available for all species of western Eurasian oaks (*Quercus*; Denk and Grimm 2010, >900 sequences; complemented by Simeone et al. 2018; see Piredda et al. 2020 for first HTS data) and numerous species of Betulaceae (Forest et al. 2005). The sequence divergence (*Supplementary file S4*) observed in our beech sample largely matches that of oaks and corresponds to inter-generic divergence in Betulaceae. The divergence between 'outgroup' O-type(s) and 'ingroup' X-I-A-B lineage is higher than between genera of the same Betulaceae subfamily; it approaches the divergence found between oak subgenera and sections.

The divergence between A- and B-types exceeds inter-species differentiation in oaks (cf. Denk and Grimm, fig. 2) and matches inter-generic differences found between *Corylus* and other members of the Coryloideae (cf. Forest and Bruneau 2000, fig. 2). The species studies in this pilot study typically have large and stable population sizes, and are widespread in parts of temperate to boreal Eurasia (e.g. Peters 1997). Beeches are wind-pollinated and disperse short-range via seeds (jaybirds being probably the most effective dispersal vector; e.g. Ridley 1930; Johnson & Adkisson 1985) and including mast years with extreme seed production (see e.g. Hilton & Packham 2003 for a historical review). Within their climax climates, they are the dominant tree species including relict areas such as the Transcaucasus (see e.g. Denk, Frotzler & Davitashvili 2001 for Georgian relict forests). Thus, one cannot expect extreme levels of genetic drift and can conclude that the main 5S-IGS sequence types reflect deep splits (much) predating the formation of extant species.

5.1. Evidence for sequence degradation (pseudogeny)

In some cases, increased 5S-IGS diversity has been linked to pseudogeny. Little is known about the GC content in functional 5S repeat units and its non-coding intergenic spacers (Symonová 2019). The 5' and 3' part of the 5S rRNA gene is highly conserved and identical in *Fagus* (our data) and *Quercus* (reference accessions AJ242950, AJ242948). Increased numbers of mutations that may be detrimental for the functioning of the 5S rRNA genes appear to be confined to (very) rare variants (cf. *Supplement file S4*; ODA subfolder *4693Data*). Exceptions are 'European O', 'Relict lineage' and 'Crenata A' variants, which commonly show signs of sequence degradation, especially also in the flanking 5S rRNA genes. As a general trend, potentially pseudogenic transitions (consensual C→T, consensual G→A) are more often found in *F. japonica* O-type than in its I-type variants (note the higher spread of GC contents for type O vs. type I; **Section 4.1, Fig. S9**); a similar observation can be made for *crenata-sylvatica* lineage A- (more common) and B-types (rare; see also **Section 4.2**, **Fig. S13**).

Table S3 | Number of visibly pseudogenic sequence variants; variants identified by compilation of sequence patterns (*Supplementary file S4*) and/or showing markedly deviating flanking gene regions in bird's eye view of the block-aligned data (NEXUS-file included in ODA). NIV = Number of variants, TA = total abundance (sum of all variants); percentages refer to the total of the given lineage/type. NIV and TA represent minimum approximates.

Type/ lineage	Class (NV)	NV	TA
Japonica O	Specific japonica	14	226
European O	Ambiguous Greek orientalis-sylvatica (3) Specific crenata (2) Specific Iranian orientalis (1) Specific sylvatica (2)	8	108
Relict lineage	Ambiguous west-Eurasian (1) Ambiguous Iranian orientalis-sylvatica (2) Ambiguous Greek orientalis-sylvatica (1) Specific crenata (2) Specific Iranian orientalis (2) Specific sylvatica (5)	13	164
Unique variant ('not-I')	Specific sylvatica (1)	1	5
Type A	Ambiguous west-Eurasian (1) Ambiguous Greek orientalis-sylvatica (4) Specific crenata (8) Specific Greek orientalis (3) Specific sylvatica (1)	17	1147[a]
Type B	Specific crenata (1) Specific Iranian orientalis (3) Specific Greek orientalis (3) Specific sylvatica (5)	12	160

[a] High number due to one variant shared by Greek *F. orientalis* and *F. sylvatica* s.str., the most common "ambiguous" variant of the *crenata-sylvatica* lineage ("ASOG" in the 38-tip set; TA = 1016)

Since the HTS approach is amplicon-based, highly pseudogenic 5S rDNA arrays will not be captured, or with much-decreased efficiency. While pseudogeny can hinder phylogenetic tree inference because of the conflicting signal from pseudogenic mutations and risk of long-branch attraction, pseudogenic nuclear spacer data have proven to be highly informative regarding past reticulations and identification of parentage in stabilized allopolyploids (e.g. Hugall, Stanton & Moritz 1999; Manen 2004; Won & Renner 2005; Grimm & Denk 2008; Vierna *et al.* 2013; Volkov *et al.* 2017).

5.2. Multi-locus organisation of 5S arrays

The finding of two main sequence clusters in each sample coincides with the only available cytogenetic data of Ribeiro et al. (2011), showing *F. sylvatica* with two 5S rDNA pericentromeric loci (and four terminal 35S rDNA loci; arrays encoding for the 18S, 5.8S, 25S rDNAs). Two sequentially distinct, common and ± conserved length groups (cf. Fig. 7; Supplement file S1) can point to genomic paralogy. In contrast to (topological) paralogy as used in phylogenetic literature, paralogy in a strict (genetic) sense implies that there are two (or more) 5S arrays per haplome. In polyploids and stabilized diploids, hom(o)eology (cf. Cronn et al. 2002), i.e. more than two 5S arrays per genome, has been observed as well in many plant groups (Yang et al. 2020; Piredda et al. 2020 and references therein), as well as in bony fish (*Xenopus*: Ford & Southern 1973, Cohen et al. 1999) and elasmobranchs (Symonová 2019). Homeologues are orthologues, i.e. they have the same locus (and function) but can evolve independently due to the lack of inter-haplome, inter-array concerted evolution. As consequence, homeologues inflict the same signal conflicts as paralogues, leading to the long-known issue of topological paralogy (e.g. Sanderson & Doyle 1992; Ebach 1999; Bailey *et al.* 2003). In beech, however, the situation is more complex than usually seen in (supposedly) diploid species and expected from cytology that identified two paralogous 5S rDNA arrays. Unless being a F1-hybrid or backcross, a diploid individual is expected to have two highly similar haplomes. In the case of beech, we thus would expect exactly two, potentially very distinct groups of 5S copies.

In *F. japonica*, the longer, GC-rich and shorter GC-poorer length classes represent two highly divergent, distantly related lineages, the outgroup O-type and the I-type, close to the ingroup variants. Both are abundant; I- and O-type variants are co-dominant. The basic hypothesis would be that each type is linked to one of the two paralogous 5S rDNA loci (per haplome). Type X variants, representing a second ‘ingroup’ lineage of ambiguous phylogenetic affinity, may be rare but nevertheless exist and show no signs of sequence degradation, pointing either to a third locus, or intra-locus/intra-array mixing of 5S copies of different evolutionary sources.

The data situation parallels cloned ITS data: ITS regions of *F. japonica* and its sisters *F. engleriana* (mainland China) and *F. multinervis* (Ullung Do, Korea) show extreme length and sequence heterogeneity with up to three divergent main sequence types, while the ITS of the *F. crenata-sylvatica* lineage is more homogeneous and poorly sorted (Denk et al. 2005; Grimm et al. 2007). The high ITS divergence corresponds with the cytological results of Ribeiro *et al.* (2011), who reported four NORs (Nucleolus Organizer Regions) for the diploid *F. sylvatica.*

In the 5S-IGS of the *crenata-sylvatica* lineage, the length and sequence differences are less pronounced with the longer A-type variants being generally less derived and less abundant than the shorter much more abundant and more diverse B-type variants (Figs 3–5; Supplement file S2, S5). The samples exhibit increasing dominance of type B over type A: weakly developed in Iranian *F. orientalis*, increased in European samples following a (south)east/(north)west gradient (Greece – Italy – Germany), and strongest in *F. crenata* (type A almost absent). In addition, putative relicts (Fig. 5) can be found that lack/mix features of A- and B-type variants (Supplement file S5) or are clearly related to the outgroup variants of the *F. japonica* type O ('European O', figs 3, 4, 6; Supplement file S2). Hence, one can detect up to three major length classes referring to four principle 5S-IGS types of disparate phylogenetic affinity.

The divergence between the *F. japonica* ingroup (type I) and outgroup (type O) variants would well fit with paralogy, independently evolving 5S rDNA array loci. Notably, although variants related to or with stronger affinity to the 'Japonica O'-type are occasionally found in the *crenata-sylvatica* lineage, they are extremely rare (cf. Table 2). A-type and 'Original B' variants have been nearly eliminated in *F. crenata* and were replaced by three sequentially distinct, and *crenata*-specific B-types, mainly 'Crenata B2' (Supplement file S2). Intragenomic silencing of paralogues/homeologues leading to pseudogeny (reviewed in Volkov et al. 2007; see e.g. Volkov et al. 2017 for a case of an ancient allopolyploid) may cause the observed detection differences. The HTS primers bind to the highly conserved 5S rDNA. If these are strongly degraded, the intergenic spacers of such arrays will not be in our sample.

While two 5S loci would facilitate retaining an ancient polymorphism (e.g. polyploidisation, hybridisation), O- vs X/I/A/B-types, they cannot explain the diversity of the latter in the *crenata-sylvatica* lineage. In the light of the findings of Ribeiro et al. (2011), there are hence two possible scenarios: *(1)* the 'O' locus (or loci) is silenced in the *crenata-sylvatica* lineage; or *(2)* the A-/B-types replaced O-like arrays in both loci. Similarly, A-type arrays (or loci) were silenced or overprinted by B-types in *F. crenata* but not (yet) in its western Eurasian sisters. Based on the high structural and sequence diversity, counterbalanced by the large number of

identical sequences detected in each sample, a combined effect of both concerted and birth-and-death evolution models must therefore be assumed for the 5S rRNA genes in beech (Nei and Rooney 2005, Galian et al. 2014). Even if there are two (or more) loci in all species of *Fagus* (which are diploid, as far as studied), they may not be paralogous in a strict sense but act like homeologues, i.e. although they differ in position they do not differ in function and are affected by inter-array recombination and limited concerted evolution.[9] To date, little is known about the main features indicating 5S gene activity in non-model plants (e.g., copy number, GC content, secondary structure, promoter and terminator characteristics; Tynkevich and Volkov 2019).

With two 5S loci (and four NORs), *Fagus* is unique within Fagaceae. The 27 *Quercus*, seven *Castanopsis*, four *Lithocarpus*, four *Castanea* and one *Trigonobalanus* species investigated so far by fluorescent in-situ hybridisation (FISH) largely showed a single pericentromeric 5S rDNA locus (www.plantrdnadatabase.com; accessed 15/08/2020, Chokchaichamnankit and Anamthawat-Jonsson 2015). Only single individuals showed an additional locus (*Castanea mollissima*) or odd numbers of (unpaired) loci (*Lithocarpus vestitus*, *Quercus suber*), as a likely result of inter-specific hybridisation or autopolyploidisation (Chokchaichamnankit et al. 2008, Ribeiro et al. 2011). No comprehensive data are currently available for a comparison with other families within Fagales with the exception of *Corylus*, Betulaceae, showing a single 5S locus and a much lower intra-individual, intra-specific divergence (Forest and Bruneau 2000).

6. List of Included Appendices

Appendix A | Summary of *Supplementary file S2.* Amplicon GC content and length ranges for main phylogenetic types; total abundance and relative proportion.

Appendix B | Violin and scatterplots of amplicon GC content and length; sorted by samples and major types.

[9] That we did not observe recombinant (chimeric) variants may be a sampling artefact. First, the pre-processing eliminates chimeric sequences. Second, 5S rDNA monomers are unlikely to show high proportions of chimeras because, in contrast to the ITS region, the non-transcribed intergenic spacer of the 5S rDNA has no sequentially highly conserved sequence patterns facilitating cross-over. In ITS region, there are several target regions: the sequentially conserved ITS1 cleavage site, the extremely conserved 5.8S rDNA, and parts of the structurally conserved ITS2, and recombinant clones can be obtained from deep-sampled, divergent, polymorphic ITS data as in the case of *Acer*.
Grimm, G.W., Denk, T. & Hemleben, V. (2007b) Evolutionary history and systematic of *Acer* section *Acer* - a case study of low-level phylogenetics. *Plant Systematics and Evolution*, **267,** 215-253.
Grimm, G.W. & Denk, T. (2014) The Colchic region as refuge for relict tree lineages: cryptic speciation in field maples. *Turkish Journal of Botany*, **38,** 1050–1066.

7. References

Bailey, C.D., Carr, T.G., Harris, S.A. & Hughes, C.E. (2003) Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Molecular Phylogenetics and Evolution,* **29,** 435-455.

Denk, T., Frotzler, N. & Davitashvili, N. (2001) Vegetational patterns and distribution of relict taxa in humid temperate forests and wetlands of Georgia (Transcaucasia). *Biological Journal of the Linnean Society,* **72,** 287-332.

Denk, T., Grimm, G., Stögerer, K., Langer, M. & Hemleben, V. (2002) The evolutionary history of *Fagus* in western Eurasia: Evidence from genes, morphology and the fossil record. *Plant Systematics and Evolution,* **232,** 213-236.

Denk, T. & Grimm, G.W. (2009) The biogeographic history of beech trees. *Review of Palaeobotany and Palynology,* **158,** 83–100.

Denk, T. & Grimm, G.W. (2010) The oaks of western Eurasia: traditional classifications and evidence from two nuclear markers. *Taxon,* **59,** 351–366.

Denk, T., Grimm, G.W. & Hemleben, V. (2005) Patterns of molecular and morphological differentiation in *Fagus*: implications for phylogeny. *American Journal of Botany,* **92,** 1006–1016.

Ebach, M.C. (1999) Paralogy and the centre of origin concept. *Cladistics,* **15,** 387-391.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics,* **27,** 2194–2200.

Forest, F. & Bruneau, A. (2000) Phylogenetic analysis, organization, and molecular evolution of the non-transcribed spacer of 5S ribosomal RNA genes in *Corylus* (Betulaceae). *International Journal of Plant Sciences,* **161,** 793-806.

Forest, F., Savolainen, V., Chase, M.W., Lupia, R., Bruneau, A. & Crane, P.R. (2005) Teasing apart molecular- versus fossil-based error estimates when dating phylogenetic trees: a case study in the birch family (Betulaceae). *Systematic Botany,* **30,** 118–133.

Fulneček, J., Lim, K.Y., Leitch, A.R., Kovarík, A. & Matyásek, R. (2002) Evolution and structure of 5S rDNA loci in allotetraploid *Nicotiana tabacum* and its putative parental species. *Heredity,* **88,** 19–25.

Garcia, S. & Kovařik, A. (2013) Dancing together and separate again: gymnosperms exhibit frequent changes of fundamental 5S and 35S rRNA gene (rDNA) organization. *Heredity,* **111,** 23–33.

Garcia, S., Wendel, J.F., Borowska-Zuchowska, N., Aïnouche, M., Kuderova, A. & Kovarik, A. (2020) The utility of graph clustering of 5S ribosomal DNA homoeologs in plant allopolyploids, homoploid hybrids, and cryptic introgressants. *Frontiers in Plant Science,* **11,** 41 [e-pub].

Grimm, G. (2020) Fagaceae collection. figshare.

Grimm, G.W. & Denk, T. (2008) ITS evolution in *Platanus*: homoeologues, pseudogenes, and ancient hybridization. *Annals of Botany,* **101,** 403-419.

Grimm, G.W. & Denk, T. (2010) The reticulate origin of modern plane trees (*Platanus*, Platanaceae) - a nuclear marker puzzle. *Taxon,* **59,** 134-147.

Grimm, G.W. & Denk, T. (2014) The Colchic region as refuge for relict tree lineages: cryptic speciation in field maples. *Turkish Journal of Botany,* **38,** 1050–1066.

Grimm, G.W., Denk, T. & Hemleben, V. (2007a) Coding of intraspecific nucleotide polymorphisms: a tool to resolve reticulate evolutionary relationships in the ITS of beech trees (*Fagus* L., Fagaceae). *Systematics and Biodiversity,* **5,** 291-309.

Grimm, G.W., Denk, T. & Hemleben, V. (2007b) Evolutionary history and systematic of *Acer* section *Acer* - a case study of low-level phylogenetics. *Plant Systematics and Evolution,* **267,** 215-253.

Grímsson, F., Grimm, G.W., Zetter, R. & Denk, T. (2016) Cretaceous and Paleogene Fagaceae from North America and Greenland: evidence for a Late Cretaceous split between *Fagus* and the remaining Fagaceae. *Acta Palaeobotanica,* **56,** 247–305.

Hilton, G.M. & Packham, J.R. (2003) Variation in the masting of common beech (*Fagus sylvatica* L.) in northern Europe over two centuries (1800–2001). *Forestry,* **76,** 319–328.

Hugall, A., Stanton, J. & Moritz, C. (1999) Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic *Meloidogyne*. *Molecular Biology and Evolution,* **16,** 157-164.

Johnson, W.C. & Adkisson, C.S. (1985) Dispersal of beech nuts by Blue Jays in fragmented landscapes. *The American Midland Naturalist,* **113,** 319–324.

Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution,* **35,** 1547–1549.

Lamb, P.D., Hunter, E., Pinnegar, J.K., Creer, S., Davies, R.G. & Taylor, M.I. (2019) How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology,* **28,** 420–430.

Li, R.-Q., Chen, Z.-D., Lu, A.-M., Soltis, D.E., Soltis, P.S. & Manos, P.S. (2004) Phylogenetic relationships in Fagales based on DNA sequences from three genomes. *International Journal of Plant Sciences,* **165,** 311-324.

Maddison, W.P. & Maddison, D.R. (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75.

Manen, J.-F. (2004) Are both sympatric species *Ilex perado* and *Ilex canariensis* secretly hybridizing? Indication from nuclear markers collected in Tenerife. *BMC Evolutionary Biology,* **4,** 46.

Mlinarec, J., Franjević, D., Bočkor, L. & Besendorfer, V. (2016) Diverse evolutionary pathways shaped 5S rDNA of species of tribe Anemoneae (Ranunculaceae) and reveal phylogenetic signal. *Botanical Journal of the Linnéan Society,* **182,** 80–99.

Negi, M.S., Rajagopal, J., Chauhan, N., Cronn, R. & Lakshmikumaran, M. (2002) Length and sequence heterogeneity in 5S rDNA of *Populus deltoides*. *Genome,* **45,** 1181–1188.

Peters, R. (1997) Beech forests. *Geobotany,* **24,** 1-169.

Piredda, R., Grimm, G.W., Schulze, E.-D., Denk, T. & Simeone, M.C. (2020) High-throughput sequencing of 5S-IGS in oaks: Exploring intragenomic variation and algorithms to recognize target species in pure and mixed samples. *Molecular Ecology Resources*.

Renner, S.S., Grimm, G.W., Kapli, P. & Denk, T. (2016) Species relationships and divergence times in beeches: New insights from the inclusion of 53 young and old fossils in a birth-death clock model. *Philosophical Transactions of the Royal Society B.*

Ribeiro, T., Loureiro, J., Santos, C. & Morais-Cecílio, L. (2011) Evolution of rDNA FISH patterns in the Fagaceae. *Tree Genetics and Genomes,* **7,** 1113–1122.

Ridley, H.N. (1930) *The dispersal of plants throughout the world.* L. Reeve & Co. Ltd, Ashford.

Sanderson, M.J. & Doyle, J.J. (1992) Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. *Systematic Biology,* **41,** 4-17.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., J., V.H.D. & Weber, C.F. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology,* **75,** 7537–7541.

Simeone, M.C., Cardoni, S., Piredda, R., Imperatori, F., Avishai, M., Grimm, G.W. & Denk, T. (2018) Comparative systematics and phylogeography of Quercus Section Cerris in

western Eurasia: inferences from plastid and nuclear DNA variation. *PeerJ,* **6,** e5793 [e-pub].

Symonová, R. (2019) Integrative rDNAomics – Importance of the oldest repetitive fraction of the eukaryote genome. *Genes,* **10,** 345 [e-pub].

Tukey, J. (1949) Comparing individual means in the analysis of variance. *Biometrics,* **5,** 99–114.

Tynkevich, Y.O. & Volkov, R.A. (2019) 5S ribosomal DNA of distantly related *Quercus* species: Molecular organization and taxonomic application. *Cytologia Genetica,* **53,** 459–466.

Vierna, J., Wehner, S., Höner zu Siederdissen, C., Martínez-Lage, A. & M., M. (2013) Systematic analysis and evolution of 5S ribosomal DNA in metazoans. *Heredity*.

Volkov, R.A., Panchuk, I.I., Borisjuk, N.V., Hosiawa-Baranska, M., Maluszynska, J. & Hemleben, V. (2017) Evolutional dynamics of 45S and 5S ribosomal DNA in ancient allohexaploid *Atropa belladonna*. *BMC Plant Biology,* **17,** 21 [e-pub].

Won, H. & Renner, S.S. (2005) The internal transcribed spacer of nuclear ribosomal DNA in the gymnosperm Gnetum. *Molecular Phylogenetics and Evolution,* **36,** 581-597.

Appendix A. General structural features (amplicon length and GC content), number of non-identical variants (NIV), and (proportional) total abundance (TA/PA) of main types per sample**.** Dominant and co-dominant (regarding both diversity and abundance) types highlighted by grey shading.

							Shared[b]		Specific					
		GC content [%]	Length [bp]		Sum[a]		'Original A/B'		'Iranian B1'		'Crenata B'		'European A/B'	
Sample	**Type**	Median Range	Median Range	**NIV**	**TA**	**PA**	**NIV**	**PA**	**NIV**	**PA**	**NIV**	**PA**	**NIV**	**PA**
06 – *Fagus japonica*	O	40.9 38.6–42.5(–44.3)	306 262–307	**398**	9593	**30%**	*Japonica*- specific							
	Short O	43.6 41.7–45.1	204 203–204[e]	97	1854	6%	*Japonica*- specific							
	I	38.6 36.1–40.8	262 197–264	**386**	20116	**62%**	*Japonica*- specific							
	X	38.1 35.9–38.9	260 258–262	26	662	2%	*Japonica*- specific							
05 – *F. crenata*	O	– 38.2; 41.0; 42.6	– 266; 304	3	30	0.1%	Sister lineage of 'Japonica O'							
	A	36.5 33.2–38.8	274 259–283(296)	31	199	0.7%	29	0.7%	–	–	–	–	0	0%
	B[c]	37.8 33.8–39.9	261 252–272	1215	27098	99%	10	16.8%	1	0.03%	**1204**	**82%**	0	0%
	Relict lin.	– 35.1–39.6	– 259–268	5	48	0.2%	Rare, derelict type of the *crenata-sylvatica* lineage							
04 – Iranian	O	– 39.7; 40.1; 42.3	– 307; 304; 260	3	24	0.1%	Sister lineage of 'Japonica O'							
F. orientalis	A	36.6 34.8–38.4	269 258–279	388	8466	42%	**386**	**42%**	–	–	–	–	0	0%
	B	37.4 35.3–39.4	256 247–269	310	11706	58%	**296**	**52%**	13	5%	1	0.005%	0	0%
	Relict lin.	– 36.4–39.8	– 256–269	5	31	0.2%	Rare, derelict type of the crenata-sylvatica lineage							
11 – Greek	O	– 37.5–41.7	– (266)303–304	8	26	0.1%	Sister lineage of 'Japonica O'							
F. orientalis	A	36.8 34.9–39.6	273 258–279	235	4932	27%	30	2%	–	–	–	–	**216**	**25%**
	Short A	39.8 38.6–41.0	166 166 [constant]	17	978	5%	0	0%	–	–	–	–	17	5%
	B[d]	37.9 33.6–41.1	261 (177–)231–267	334	12546	68%	149	23%	2	0.1%	1	0.03%	**180**	**45%**
	Relict lin.	– 36.9–40.0	– 256–268	4	25	0.1%	Rare, derelict type of the crenata-sylvatica lineage							
14 – *F. sylvatica* s.str.	O	– 37.5–40.1(41.7)	– 303–304(266)	9	33	0.3%	Sister lineage of 'Japonica O'							
(Italy)	A	36.6 34.4–39.6	274 258–280	241	4142	38%	10	1%	–	–	–	–	**228**	**37%**
	Short A	– 40.4	– 166	2	12	0.1%	0	0%	–	–	–	–	2	0.1%
	B	38.3 35.7–41.5	261 212–262	312	6605	61%	57	12%	1	0.04%	0	0%	**254**	**49%**
	Relict lin.	– 35.8–39.8	– 256–268	10	28	0.3%	Rare, derelict type of the crenata-sylvatica lineage							
12 – *F. sylvatica* s.str.	O	– 37.5–40.1(41.7)	– 303–304(266)	8	63	0.2%	Sister lineage of 'Japonica O'							
(Germany)	Not I[e]	– 36.0	– 253	1	5	0.01%	Ingroup relict type of unclear affinity							
	A	36.6 34.7–39.1	274 258–280	308	6176	17%	7	0.1%	–	–	–	–	**298**	17%
	B	38.5 35.4–40.8	261 243–262	844	30142	83%	18	0.6%	5	0.1%	0	0%	**820**	**82%**
	Relict lin.	– 35.8–39.8	– 256–268	11	90	0.2%	Rare, derelict type of the crenata-sylvatica lineage							

[a] Sum includes variants placed at the root branch of the respective clades

[b] Far the most 'Oriental A' types are specific for Iranian *F. orientalis* (see Figs 2, 3)

[c] Most variants are of type 'Crenata B2' (total abundance = 21609)

[d] Includes one variant with an abundance = 9 (GC content = 38.1%; length = 260) placed at the I-B-clade root.

[e] See *Supplementary file S4* ("SyG3")

[f] Corrected for wrongly clipped 5' ends found in four 'short O' variants (see Supplementary files S1, S2)

Appendix B | Violin and scatterplots including all 5S-IGS types