Complete pan-plastome sequences enable high resolution phylogenetic classification of sugar beet and closely related crop wild relatives

Katharina Sielemann[1,2], Boas Pucker[1,3,4], Nicola Schmidt[5], Prisca Viehöver[1], Bernd Weisshaar[1], Tony Heitkam[5*], Daniela Holtgräwe[1*]

[1]Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec) & Faculty of Biology, Bielefeld University, 33615 Bielefeld, Germany

[2]Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Bielefeld University, 33615 Bielefeld, Germany

[3]Evolution and Diversity, Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom

[4]Institute of Plant Biology, TU Braunschweig, Braunschweig, Germany

[5]Faculty of Biology, Institute of Botany, Technische Universität Dresden, 01069 Dresden, Germany

*** Correspondence (These authors share senior authorship.):**
Tony Heitkam, Daniela Holtgräwe
tony.heitkam@tu-dresden.de, dholtgra@cebitec.uni-bielefeld.de

Keywords: sugar beet, *Beta vulgaris*, *Beta*, *Corollinae*, *Patellifolia*, chloroplast, plastome assembly, phylogeny, phylogenomics

Short title: Pan-plastome of cultivated and wild beets

Abstract

Background

As the major source of sugar in moderate climates, sugar-producing beets (*Beta vulgaris* subsp. *vulgaris*) have a high economic value. However, the low genetic diversity within cultivated beets requires introduction of new traits, for example to increase their tolerance and resistance attributes – traits that often reside in the crop wild relatives. For this, genetic information of wild beet relatives and their phylogenetic placements to each other are crucial. To answer this need, we sequenced and assembled the complete plastome sequences from a broad species spectrum across the beet genera *Beta* and *Patellifolia*, both embedded in the Betoideae (order Caryophyllales). This pan-plastome dataset was then used to determine the wild beet phylogeny in high-resolution.

Results

We sequenced the plastomes of 18 closely related accessions representing 11 species of the Betoideae subfamily and provide high-quality plastome assemblies which represent an important resource for further studies of beet wild relatives and the diverse plant order Caryophyllales. Their assembly sizes range from 149,723 bp (*Beta vulgaris* subsp. *vulgaris*) to 152,816 bp (*Beta nana*), with most variability in the intergenic sequences. Combining plastome-derived phylogenies with read-based treatments based on mitochondrial information, we were able to suggest a unified and highly confident phylogenetic placement of the investigated Betoideae species.

Our results show that the genus *Beta* can be divided into the two clearly separated sections *Beta* and *Corollinae*. Our analysis confirms the affiliation of *B. nana* with the other *Corollinae* species, and we argue against a separate placement in the *Nanae* section. Within the *Patellifolia* genus, the two diploid species *Patellifolia procumbens* and *Patellifolia webbiana* are, regarding the plastome sequences, genetically more similar to each other than to the tetraploid *Patellifolia patellaris*. Nevertheless, all three *Patellifolia* species are clearly separated.

Conclusion

In conclusion, our wild beet plastome assemblies represent a new resource to understand the molecular base of the beet germplasm. Despite large differences on the phenotypic

level, our pan-plastome dataset is highly conserved. For the first time in beets, our whole plastome sequences overcome the low sequence variation in individual genes and provide the molecular backbone for highly resolved beet phylogenomics. Hence, our plastome sequencing strategy can also guide genomic approaches to unravel other closely related taxa.

Background

As the crop plant *Beta vulgaris* (sugar beet) has a high economic value (1), continuous crop development is essential to enhance stress tolerances and resistances against pathogens. The White Silesian Beet provided the germplasm for sugar beet (2) and is a derivative of wild sea beet (*B. vulgaris subsp. maritima*). This leaves, similar to the situation in many domesticated crops, only a narrow genetic base for sugar beet breeding (3). Additionally, early sugar beet breeding has focused mainly on increasing yield. This caused strong domestication bottle necks and removed many useful traits that may benefit plant fitness (3,4). The higher genetic variation in crop wild relatives of sugar beet offers potential that might be harnessed to introduce desired traits. Thus, giving insight into the genomic basis of wild beets is progressively moving into the focus of beet breeding research (5,6).

Phylogenetically, wild beets belong to the Betoideae (order Caryophyllales) and are separated in the genera *Beta* and *Patellifolia*, with all cultivated beets belonging to the genus *Beta*. The genus *Beta* is then further subdivided into at least two sections, *Beta* and *Corollinae*. In general, the section *Beta* is widespread across Western Europe, whereas *Corollinae* species are generally distributed across the eastern Mediterranean area and South-West Asia (1,7) (Additional file S1A). Despite the long history of different systematic treatments (1,7–12), the phylogenetic relationships of *Beta* and *Patellifolia* species are still a matter of ongoing debate (as reviewed in (12)). Especially the subdivision of the genus *Beta* into three sections (*Beta*, *Corollinae*, and *Nanae*) is discussed, with the pending suggestion to integrate *B. nana* into the section *Corollinae*, hence disbanding the section *Nanae* (7,12,13). Similarly, as the *Beta* section *Corollinae* harbors a highly variable polyploid/hybrid complex, including di-, tri-, tetra-, penta-, and hexaploid forms, the species boundaries are far from resolved. Regarding the sister genus, there is still an ongoing discussion on whether the morphologically variable *Patellifolia* comprise three distinct species (*P.*

patellaris, *P. procumbens*, and *P. webbiana*) or only two or even one (7,9,11). Resolving the unclear wild beet relationships may inform beet improvement programs and contribute to the development of new, better equipped beets.

The plastome is well-suited for the reconstruction of phylogenies due to high structural conservation, a conserved evolutionary rate, uniparental inheritance, and high abundance of DNA across all species (14,15). Historically, systematic information was obtained from plastome sequence restriction site variants, inversions, single nucleotide variants (SNVs), or spacers in single genes. Although this has led to a range of wild beet phylogenies resolving relationships on the level of genera and sections, these are often based on only a few species and contain collapsed branches due to low genetic variation. In contrast, the investigation of whole plastome sequences may enhance the resolution of phylogenetic relationships (14,16–18). As gene sequences and intergenic regions can be included and combined, whole plastome sequence analyses enable the detection of well-supported phylogenetic relations on the species- and even on the accession level. Thus, plastid genomics may offer a route to clarify many of the pending questions regarding the wild beet phylogeny.

The plastome sequence of most angiosperms comprises a total of 79 protein-coding genes, 4 rRNA genes, and 30 tRNA genes (19). The quadripartite structure is characteristic for plastome sequences comprising a large (LSCR) and a small single-copy region (SSCR) as well as two inverted repeats (IRs) (20,21), all contributing to a total length of 120 kb to 210 kb (20). This difference in size can be mainly attributed to the IRs that range from 6 kb to 76 kb in length (21–23). The relative orientation of the SSCR between the IRs differentiates two structural variants which occur simultaneously in a single cell and might have been previously mistakenly annotated as differences between species (24).

The Caryophyllales, including *B. vulgaris*, contain canonical plastomes, harboring all hallmarks typical for angiosperm plastome sequences as described above (25,26). For the wild beet species, until now, no plastome assembly is available, and of our investigated species, only the plastome sequence of *B. vulgaris* subsp. *vulgaris* was published previously (27). A detailed, plastome- and mitogenome-based evolutionary positioning of species outside of the section *Beta* is still missing but needed to answer some of the unresolved issues in beet systematics.

Here, we resolve the phylogenetic relationships within the Betoideae at high resolution through genome-wide comparison based on complete plastome assemblies and reads from both, the plastome and the mitogenome. Eleven different *Beta* and *Patellifolia* members, spanning the previously neglected plastome sequences of the *Corollinae* section and the *Patellifolia* genus, are included in our analyses. For this, whole plastome sequences of up to two accessions per species are sequenced, assembled, and compared. This novel contribution to the Betoideae pan-plastome intends to clarify the phylogenetic relationships of wild beets on a species-level and provides an important resource for further studies of beet wild relatives.

Results

Our pan-plastome dataset comprises 18 different accessions, including a biological replicate of *B. corolliflora*, which leads to 19 plastome assemblies in total (see Methods, Table 2). To provide a basis for comparative plastome analysis, all plastome sequences were fully assembled. Out of those, 17 were split into three scaffolds (LSCR, SSCR, and IR), apart from Bmar1 (four scaffolds) and Bnan2 (six scaffolds). Collapsed IR regions were confidently identified in all plastome assemblies based on a doubled average read coverage in comparison to the single copy regions as well as a gene content that is characteristic and expected for the IR region. Average read coverage and assembly length are shown in Table 1. The distribution of these values is shown in Additional file S1B. Circular and linear plots of a representative selection of plastome assembly sequences are provided in Additional file S1C.

Table 1: Plastome assembly statistics. Average read coverage values and assembly lengths (in bp) for each region and the complete assemblies are shown. Abbreviations: LSCR = Long single copy region; SSCR = Short single copy region; IR = Inverted repeats.

Total coverage	**LSCR coverage**	**SSCR coverage**	**IR coverage**	**Total length**	**LSCR length**	**SSCR length**	**IR length**

396 ± 73	261 ± 81	265 ± 98	662 ± 100	150,519 ± 892	83,496 ± 621	17,845 ± 257	24,588 ± 307

Comparing the plastome assemblies of all *Beta* vs. *Patellifolia* species, the average length of the four *Patellifolia* plastome sequences (avg. 151,621 bp) is higher than for the 15 *Beta* plastome sequences (avg. 150,225 bp). This length difference can be mainly assigned to the LSCR (avg. *Beta* 83,401 bp; avg. *Patellifolia* 83,853 bp) and to the IRs (avg. *Beta* 24,435 bp; avg. *Patellifolia* 25,166 bp). However, the SSCR is longer in *Beta* plastome sequences (avg. 17,954 bp) when compared to the plastome sequences of all *Patellifolia* accessions (avg. 17,437 bp).

Interestingly, plastome assemblies of *B.* section *Corollinae* (avg. 36.67 %) show a higher GC content when compared to *B.* section *Beta* plastome sequences (avg. 35.81 %). The total length of *B.* section *Corollinae* plastome sequences (avg. 150,504 bp from nine species) is higher in comparison to *B.* section *Beta* plastome assemblies (avg. 149,808 bp from six species). This length difference is visible for all regions of the plastome sequence (LSCR, SSCR and IRs).

The final plastome assemblies were subsequently annotated and the alignment identity of all regions included in the phylogenetic analysis was assessed for gene regions and intergenic regions, respectively (Figure 1). The alignment identity is significantly higher for gene sequences when compared to intergenic regions. This significant difference was obtained when amaranth, quinoa, and spinach were included as outgroups (Additional file S2A) (avg. gene/intergenic regions 90.73/83.55 %; Mann-Whitney-U test; p 2e-10) as well as without outgroup reference sequences (Additional file S2B) (avg. gene/intergenic regions 97.26/94.93 %; Mann-Whitney-U test; p ≈ 1e-09). *Rrn* genes show high similarity among all plastome genes, whereas *ycf1* and *rpl22* show the greatest variance between all investigated accessions. The intergenic region between the genes *ycf4* and *cema* contributes most to the differences in the alignment.

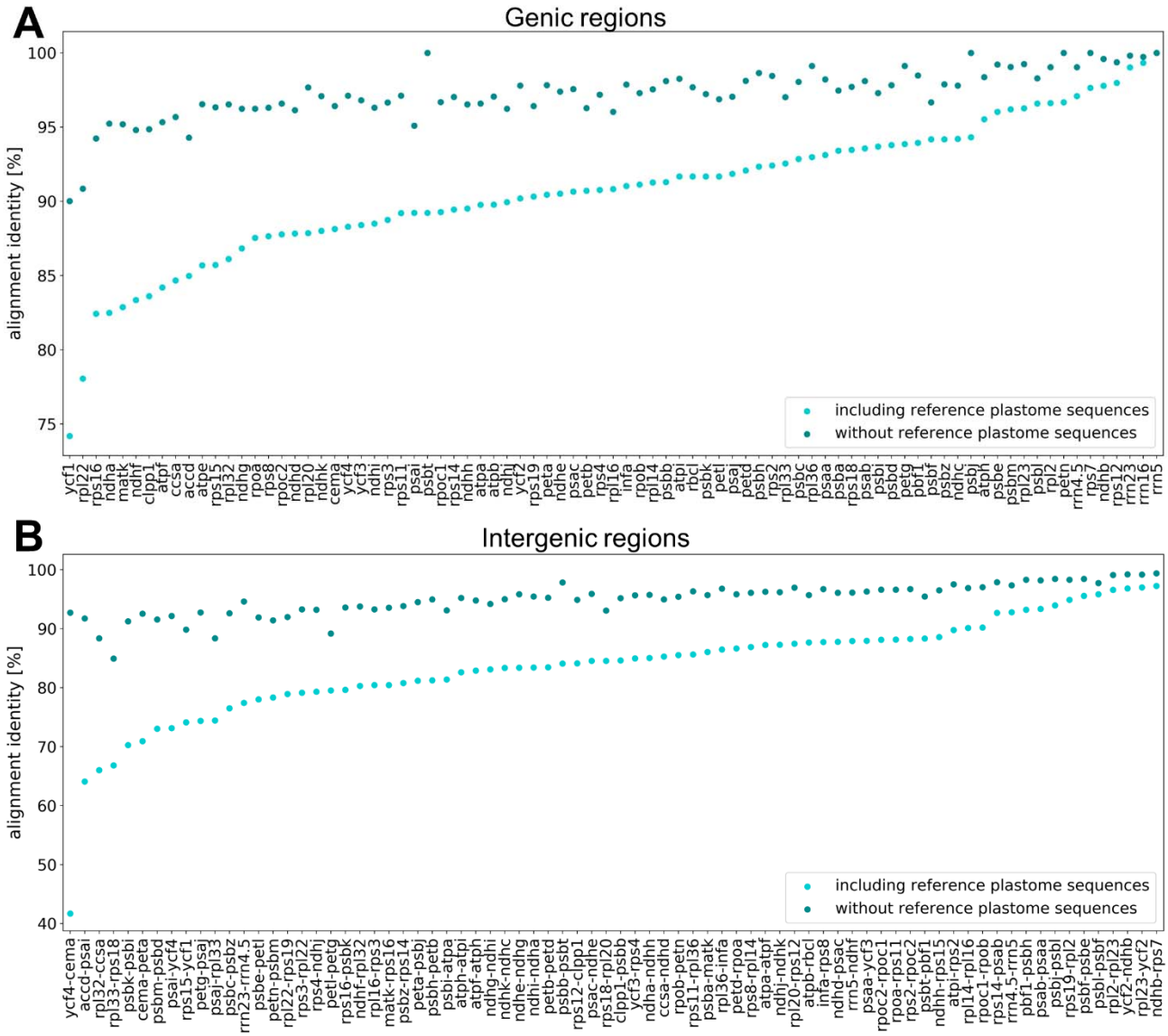


Figure 1: Alignment identities of gene sequences (A) and intergenic regions (B). The sequences (x-axis) are ordered based on the alignment identity represented by the y-axis. The alignment identity including amaranth, quinoa, and spinach as outgroup (light blue) as well as the alignment identity without the reference plastome sequences (dark blue) are shown.

The distributions of SNVs (Figure 2A) and InDels (Figure 2B) throughout the plastome sequences were further investigated. InDels are mostly absent from gene regions and SNVs are in general more frequent than InDels. Further, some clear hotspots of SNVs and

InDels can be detected in the intergenic regions *psbk-psbi*, *ycf4-cema*, *rpl33-rps18*, *psaj-rpl33* and *rpl32-ccsa*.

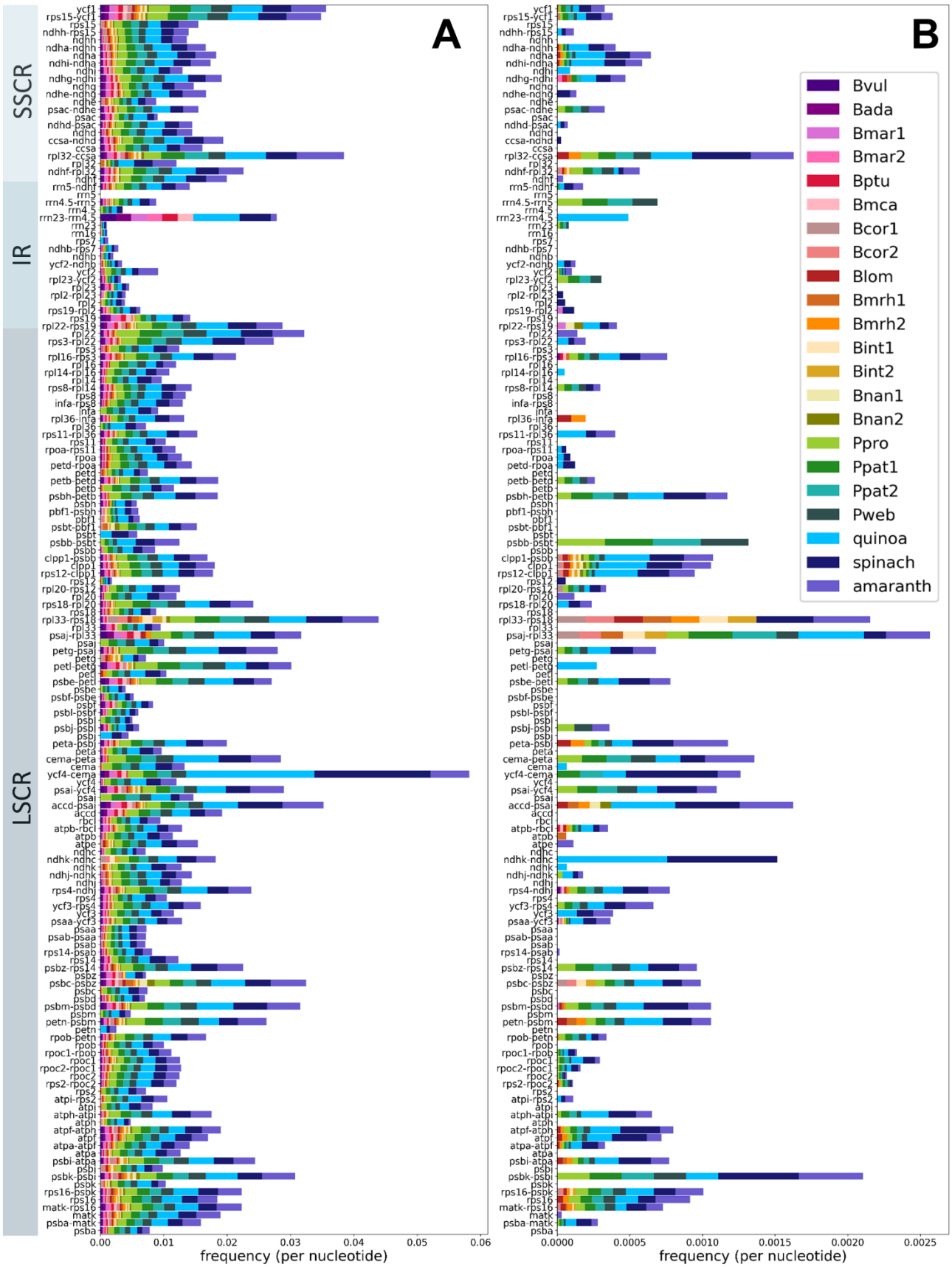

A
B
SSCR
IR
LSCR
Bvul
Bada
Bmar1
Bmar2
Bptu
Bmca
Bcor1
Bcor2
Blom
Bmrh1
Bmrh2
Bint1
Bint2
Bnan1
Bnan2
Ppro
Ppat1
Ppat2
Pweb
quinoa
spinach
amaranth
0.00
0.01
0.02
0.03
0.04
0.05
0.06
frequency (per nucleotide)
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
frequency (per nucleotide)

Figure 2: Number of SNVs and InDels in the assembled plastome sequences. Number of SNVs (A) and number of InDels (B) in the sequence alignments normalised by the length of the respective gene sequence/intergenic region and by the number of species/accessions. The gene names and intergenic regions on the x-axis are ordered based on the arrangement in the plastome assembly. Amaranth, quinoa, and spinach were included as outgroup.

As InDels with a length, which is a multiple of three, do not influence the reading frame (28), we expected that the proportion of these InDels (which are multiples of three) is higher in gene regions compared to the proportion in intergenic regions. Indeed, we observed that 43.4 % of the InDels in gene regions were a multiple of three, whereas this applies to only 29.6 % of the InDels in intergenic regions (Fisher's exact test; p ≈ 3e-8; Figure 3 [arrows]).

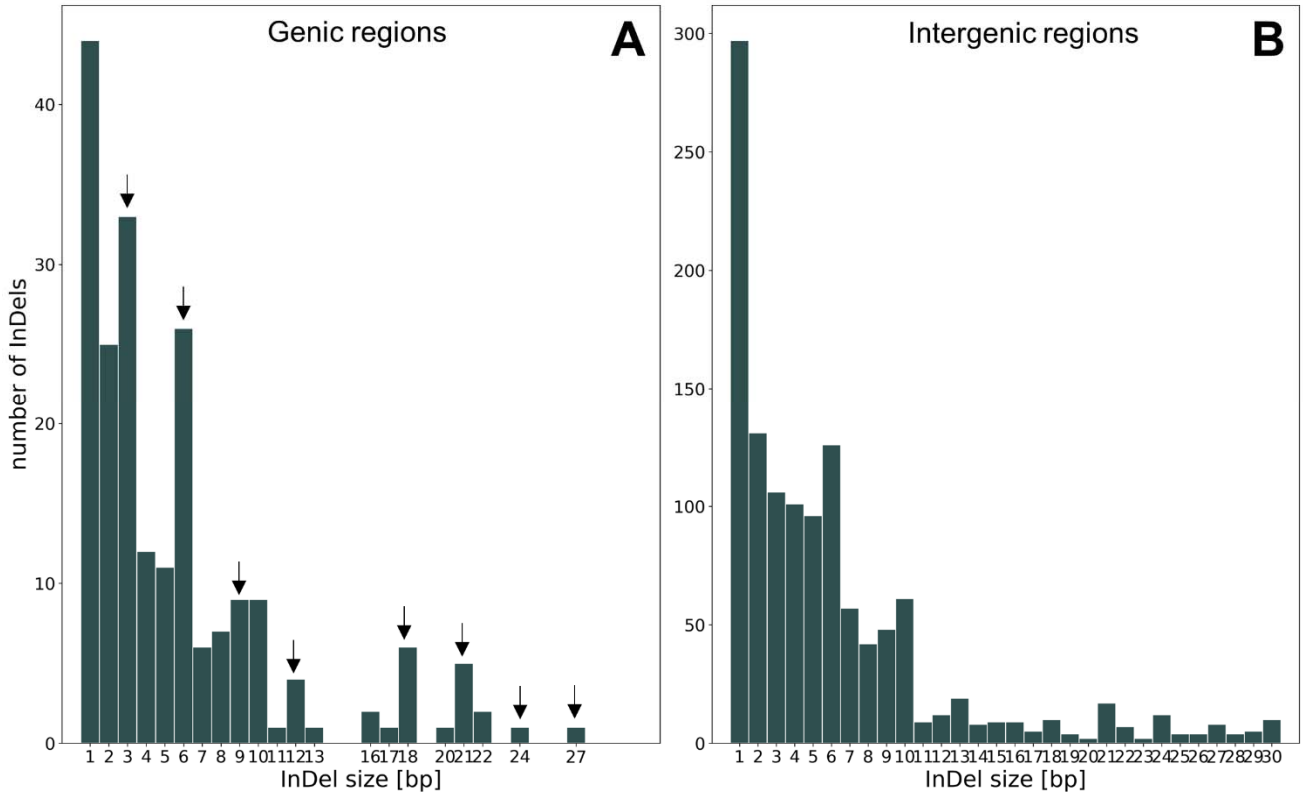
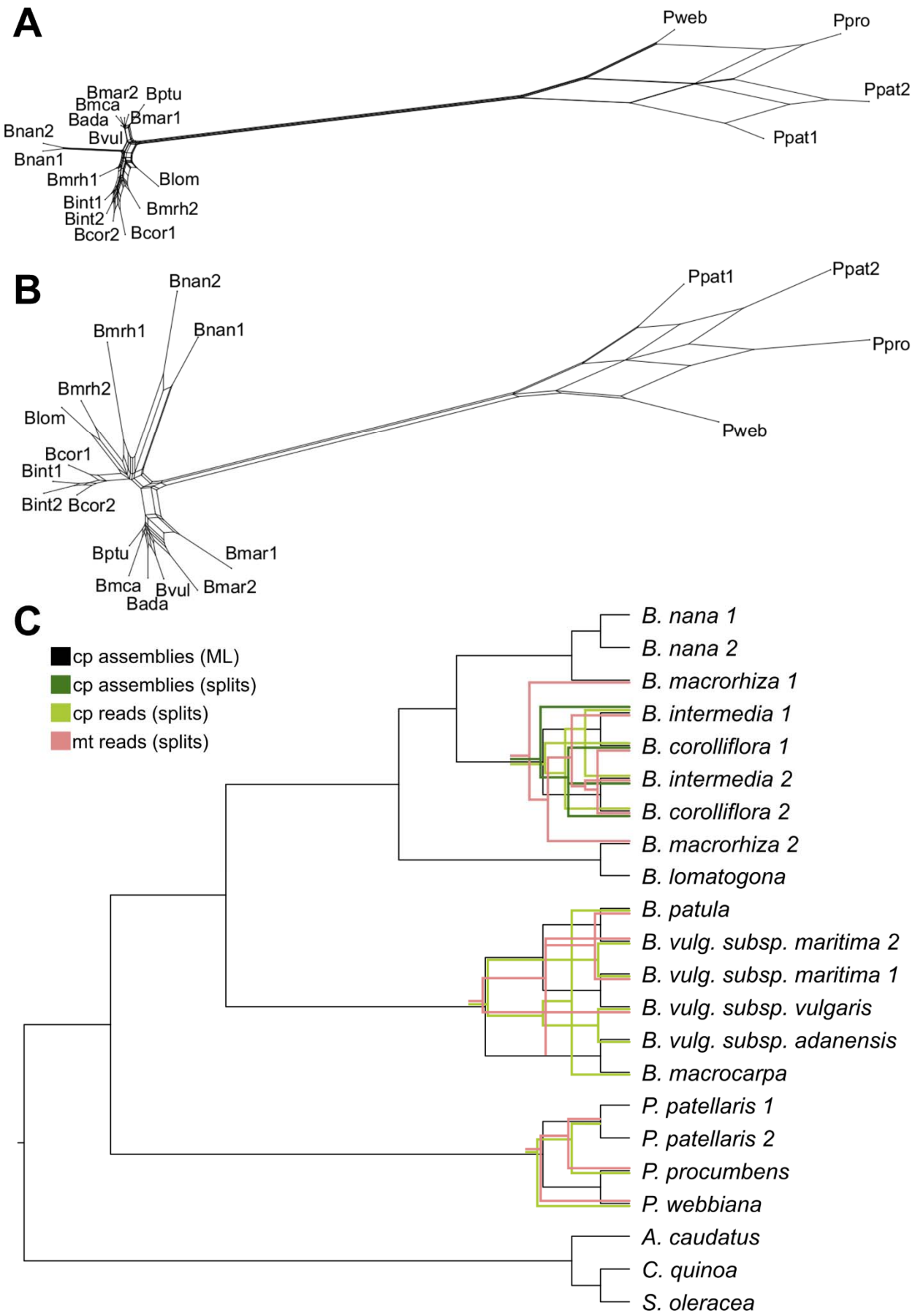


Figure 3: Number of InDels based on size in gene sequences (A) and intergenic regions (B). The arrows represent InDels with a size of a multiple of three. Please mind the variable y-axis.

Phylogenetic relations of the Betoideae subfamily were inferred from colored de Bruijn graph-based splits (Figure 4) as well as by an alignment-based maximum likelihood (ML) analysis (Figure 5). Mitochondrial and plastid read derived kmers were used to calculate phylogenetic splits, and annotated gene sequences as well as intergenic regions derived from the plastome assemblies were used for the ML analysis. A clear separation of *Patellifolia*, *B.* section *Beta* and *B.* section *Corollinae* samples is visible in all four phylogenetic trees. In comparison to the ML tree based on fully assembled plastome sequences (Figure 4C, black), the tree based on plastome assembly-derived splits (Figure 4C, dark green) shows only one difference among the *B. intermedia* and *B. corolliflora* accessions. The phylogenetic relationships derived from kmers show a few additional differences (Figure 4C, light green and pink). These differences comprise for example the assignment of the four *Patellifolia* accessions/species to a clade consisting of both *P. patellaris* accessions and a separate clade formed by *P. procumbens* and *P. webbiana* for the assembly-based phylogenies (Figure 4C, black and dark green), whereas the kmer based phylogenies show a separate clade for *P. webbiana* and a second clade comprising the other three *Patellifolia* accessions/species (Figure 4C, light green and pink). The calculation of the weighted F1 score, the weighted symmetric set distance and the Robinsons-Foulds distance shows that there is a high identity between all splitstree results (based on cp_reads, mt_reads and cp_assemblies) (Additional file S1D). The usage of different input data formats (reads vs. assemblies) has a larger impact than the usage of different datasets (chloroplasts vs. mitochondria) meaning that these splitstree results show a higher divergence.


A
Pweb
Ppro
Ppat2
Ppat1
Bmar2
Bptu
Bmca
Bada
Bmar1
Bnan2
Bvul
Bnan1
Bmrh1
Blom
Bint1
Bint2
Bmrh2
Bcor2
Bcor1
B
Bnan2
Bmrh1
Bnan1
Bmrh2
Blom
Bcor1
Bint1
Bint2
Bcor2
Bptu
Bmar1
Bmca
Bvul
Bmar2
Bada
Ppat1
Ppat2
Ppro
Pweb
C
cp assemblies (ML)
cp assemblies (splits)
cp reads (splits)
mt reads (splits)
B. nana 1
B. nana 2
B. macrorhiza 1
B. intermedia 1
B. corolliflora 1
B. intermedia 2
B. corolliflora 2
B. macrorhiza 2
B. lomatogona
B. patula
B. vulg. subsp. maritima 2
B. vulg. subsp. maritima 1
B. vulg. subsp. vulgaris
B. vulg. subsp. adanensis
B. macrocarpa
P. patellaris 1
P. patellaris 2
P. procumbens
P. webbiana
A. caudatus
C. quinoa
S. oleracea

220

Figure 4: Phylogenetic relationships of 18 different *Beta*/*Patellifolia* accessions derived from different strategies and datasets. Kmer-based trees were constructed using raw sequencing reads (A: mitochondrial reads, B: chloroplastic reads) as well as the final chloroplast assemblies as input (not shown as only used for comparison). The splitstree results (green and pink) were then compared to the ML analysis (black) (C). Discordance between the phylogenetic trees is shown in the respective color. (Abbreviations: cp=chloroplast; mt=mitochondria).

For the ML-based tree (Alignment sites / patterns: 216442 / 2441; Gaps: 0.44 %; Invariant sites: 86.73 %), the phylogenetic relationships on the species level are highly supported (high bootstrap values) (Figure 5). A phylogenetic tree based on the diagnostic set of 53 gene sequences and intergenic regions matches this phylogeny (Additional file S1E).

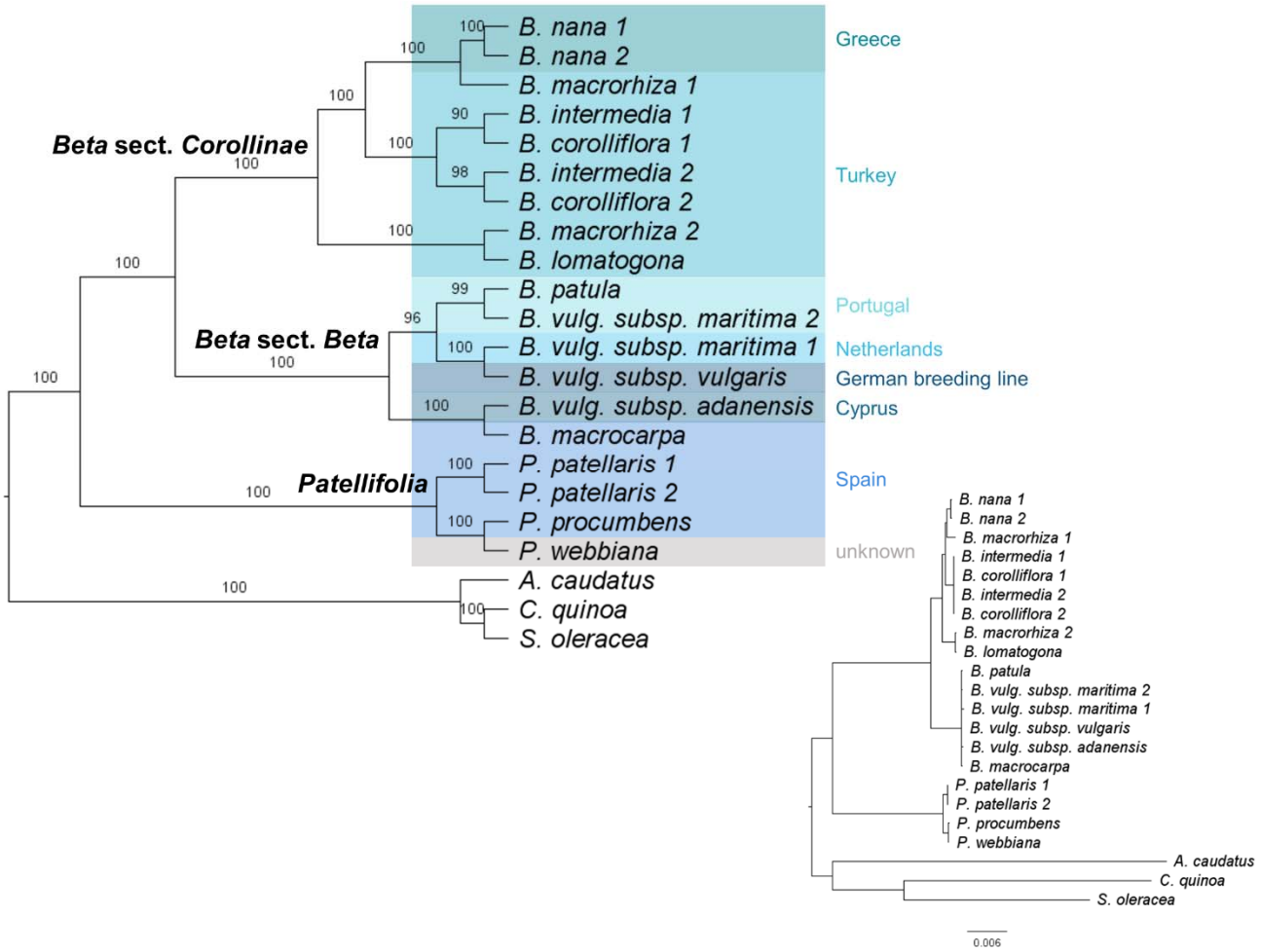


Figure 5: Plastome phylogeny for Betoideae. The tree can be divided into three groups: *B.* section *Corollinae* (8 accessions), *B.* section *Beta* (6 accessions), and *Patellifolia* (3 accessions). The plastome sequences of three Caryophyllales species (amaranth, quinoa, and spinach) were used as outgroup. Bootstrap support values are shown above each branch. The resulting phylogeny is based on the variation in 83 genes and 76 intergenic regions from the plastid genomes of 18 accessions and species (plus outgroup). Different background colors represent the sampling location and the origin of the breeding line (Bvul), respectively. Inset: Actual branch lengths based on the ML analysis.

To investigate the contribution of different regions to the phylogeny, in addition to the whole plastome sequences (genic and intergenic regions combined), sequence matrices for (I) all gene regions, (II) all intergenic regions, (III) whole coding sequences, (IV) first and second codon position and (V) third codon position only were constructed and used for the inference of phylogenetic relationships. Even though the topology of the phylogeny derived from whole coding sequences is highly similar (Additional file S1F), especially for the codon position-based matrices, alternative branches can be observed. However, substantially more nodes are only poorly supported with bootstrap values up to below 20.

The alignment of the 18s rRNA gene sequence for *B. corolliflora* (Bcor1) as representative for the *Corollinae* and *B. vulgaris subs. vulgaris* as member of the section *Beta* revealed not a single genomic difference and therefore no signal which could be used to resolve their phylogenetic relations.

The results presented here show that especially for closely related species of the same subfamily, a higher number of gene sequences and more variable intergenic regions provides greater phylogenetic resolution.

Discussion

Our proposed beet whole-plastome phylogeny is superior to single gene phylogenies

Here, we present the plastome sequence assemblies of 18 accessions covering most of the species' diversity within the beet genera *Beta* and *Patellifolia* and representing an important resource for future studies. All newly generated plastome sequences are highly similar including 79 protein coding genes and four rRNA genes distributed across a mean length of

150,519 bp (± 892bp). A previously published *B. vulgaris* subsp. *vulgaris* (KR230391) plastome sequence comprises a length of 149,722 bp (27). This is almost identical to our Bvul assembly which differs by only 1 bp in length (149,723 bp). This difference occurs in a stretch of five or six guanines in the IR region between the genes *rrn23* and *rrn16* – either a sequencing error or a biological difference. As expected, all Betoideae assemblies show high similarity as all angiosperm plastomes are highly conserved and the species investigated here are closely related, while most differences are located in intergenic regions.

Between the cultivated beet and wild beet accessions, most chloroplast genes are highly conserved, one example being *rpl2* with only a low number of polymorphisms (Figure 1) (22). The intergenic regions are significantly less conserved containing more InDels and are therefore more suitable for phylogenies on a lower taxonomic level (22). Among the beet plastomes, *ycf1* is the most informative genic region (Figure 1A), which was also detected in other plant groups, such as the Tropaeolaceae, the orchids, and the Malvales (29–31). Additionally, the *rpl22*, *matk*, *clpP1*, *ycf2*, *psaC* and *ndhF* genes were reported to be highly divergent (29,31,32), which is mainly consistent with our findings. Here, the rrn-genes, *ndhB* (already classified as gene with low divergence (29,31)), *rps12* and *rps7* can be found among the gene loci with lower variance among the investigated species.

Plastomes in general show very similar sequences with most differences occurring in non-coding regions (33). For beet and wild beets, the most informative intergenic regions are *ycf4-cema*, *accd-psai* and *rpl32-ccsa* (Figure 1B). However, for *ycf4-cema* the high difference between ‘reference’ and ‘without reference plastome-’ alignment identity should be noticed. For the Malvales, the regions *psaB-psaA*, *psbF-psbE*, *rpl2-rpl23* and *ndhH-ndhA* were identified as lowly divergent regions (31). We confirmed these for beets, except for the latter (*ndhH-ndhA*) that showed higher divergence among the wild beet plastomes. In contrast, the intergenic regions with the highest divergence in the Malvales (*ndhD-ccsA* and *rps19-rpl2*) (31) accounted for less differences among the wild beets. Nevertheless, the percent identity values among all intergenic regions are relatively similar in beets, especially when excluding the polymorphisms in the outgroup reference genomes.

As high variability among the investigated sequences is required to resolve phylogenetic relations of closely related species (34), the retrieved plastome sequences from beets and

wild beets provide an excellent resource to approach systematic treatment of the Betoideae. To further increase sequence variability, we also integrated the more diverged intergenic regions into the analysis.

In addition to the plastome-inferred ML-based phylogeny, a mitogenome- and plastome-inferred read-based phylogenetic tree was constructed based on phylogenetic splits. The resulting trees can be considered reliable due to three distinct, robust properties: (I) all splits networks show a tree-like appearance (Figure 4A, 4B), (II) the usage of different kmers leads to the same tree and (III) the usage of the geometric mean leads to the removal of samples in case no kmer occurs in the sample and both, geom and geom2 lead to the same results (with the exception of k=11, which results in a cloud-shaped network).

Differences between the phylogenetic trees derived from different strategies and datasets (Figure 4C) may be explained by chance and/or by the use of the method (all splitstree results show high identities as indicated by all comparison metrics (Additional file S1D) while differences are larger when using different strategies [reads vs assemblies] in comparison to different datasets [chloroplasts vs. mitochondria]) or by real differences in the biological nature of mitochondria and chloroplasts. Mitochondrial DNA shows a low nucleotide substitution rate when compared to chloroplast DNA (25,35). Reasons for this may be recent species hybridization or incomplete lineage sorting (36,37). Therefore, the mitogenome seems to be mostly useful at higher taxonomic levels (25) and might not be the most suitable system for the beet and wild beet accessions investigated in this study.

Compared to our pan-plastome assembly, the information derived from individual genic and intergenic regions are insufficient to fully resolve the beet phylogeny, highlighting the power of our plastome approach:

I) Investigation of specific regions of the plastome sequence (genic and intergenic regions, coding sequences, codon positions) revealed a few alternative branches for the codon position-based sequence matrix (Additional file S1F). These are marked by short internal branch lengths due to the close relationships of the species within this subfamily. Nevertheless, these conflicting relationships are only weakly supported. This is explained by the lower genetic diversity and therefore insufficient phylogenetic signal when using a smaller amount of sequences and total sequence length.

II) An approach based on high-quality single-copy nuclear genes would require a minimum coverage of about 10x (25). Moreover, nuclear genes are often part of gene families and influenced by whole genome duplication events (14). Using our available data, nrDNA sequences were selected for the phylogenetic reconstruction. Especially the ITS and ETS regions were previously used for the investigation of phylogenetic relationships, but entire nrDNA repeats (18S-ITS1-5.8S-ITS2-26/28S) were also already assembled for multiple phylogenetic studies (15,38). Unfortunately, the assembled nrDNA sequences constructed here are not useful to infer confident phylogenetic relationships as the coverage is very low (1.9x - 8.5x) and intragenomic polymorphisms of different nrDNA repeat sequences might limit the reliability of the phylogeny (15). The low bootstrap values and the low coverage make this phylogenetic tree unreliable. Therefore, we do not show these results here. Another possible explanation, apart from the low coverages, is that the biparental nature of the nuclear genome may be problematic for the inference of phylogenetic relationships (14). Previous studies already suggest that phylogenies based on nrDNA and few selected plastid sequences only weakly support relationships (30). The 18s rRNA gene sequences of representatives of the sections *Corollinae* and *Beta* are completely identical containing no phylogenetic signal to separate them.

Summarizing, our plastome-derived phylogeny benefits from the incorporation of genic and intergenic regions as well as the 'nature' of the plastome itself (as described in the Background section). Despite the low available read coverage and the low genetic diversity within our beet dataset, this leads to a highly confident phylogenetic tree. Further, the use of higher alignment lengths and the use of nucleotides instead of amino acids are favored to construct well supported phylogenies (39). Therefore, we conclude that for resolving the relationships of cultivated and wild beets, our whole-plastome-based approach is the most reliable.

Implications for the systematic placements within the Betoideae subfamily

With efforts tracing back half a century, resolving the phylogeny of the subfamily Betoideae has been already a major undertaking (1,7–10,40). However, in most cases only few selected sequence regions were targeted, leading to unresolved relations at shallower taxonomic levels or with focus on specific species or sections, i.e.:

I) In a study by Hohmann et al. (2006), the *Beta* species *B. vulgaris*, *B. corolliflora*, *B. nana* and *B. trigyna* were investigated using ITS, *trnL-trnF* spacer and *ndhF* sequences (9). Kadereit et al. (2006) provide a comprehensive analysis of a high number of different Betoideae species, finding that, *B.* section *Beta* was clearly separated from *B.* section *Corollinae*, which contained *B. nana*, *B. trigyna*, *B. macrorhiza*, *B. corolliflora* and *B. lomatogona*. As the analysis was based on ITS sequences comprising only 251 characters of which 147 were invariable, relations between species in both sections could not be resolved (7). Our study confirms the deep separation of the sections *Beta* and *Corollinae* and refines the resolution on the species level.

II) A recent comprehensive study by Romeiras et al. (2016) of phylogenetic relationships in the Betoideae is based on ITS and *matK*, *trnH-psbA*, *trnL* intron and *rbcL* sequences and investigates a high number of different species leading to the following main result: *Beta* and *Patellifolia* species are two clearly separated monophyletic groups (1). In total, three monophyletic lineages were identified: *B.* section *Beta* (*B. vulgaris* subsp. *vulgaris*, *B. vulgaris* subsp. *maritima*, *B. macrocarpa*, *B. patula*), *B.* section *Corollinae* (*B. nana*, *B. corolliflora*, *B. trigyna*) and *Patellifolia* (*P. patellifolia*, *P. procumbens* and *P. webbiana*). Exact relations within the Betoideae on a lower taxonomic level remain unclear as the branches are not well supported and collapsed. We also identify the three monophyletic groups as proposed and manage to resolve many of the previously collapsed branches.

III) Recently, Touzet et al. (2018) investigated the relationship of a wide range of *B. vulgaris* subsp. *maritima*, *B. macrocarpa* and *B. vulgaris* subsp. *adanensis* accessions based on a 3,742 bp alignment of plastome sequences and a 1,715 bp alignment of selected nuclear sequences (10). They find, based on a representative geographical sampling, that *B. macrocarpa* is a distinct lineage from the two investigated *B. vulgaris* subspecies. Despite this interesting finding, the suggested phylogeny did not focus on other important species and accessions of the Betoideae subfamily and might be further improved by the analysis of sequences with higher diversity to reach higher bootstrap values, which we achieved using intergenic and genic regions of the whole plastome sequences.

With our pan-plastome-informed datasets, we have been able to confirm many of the observations before and added an unprecedented resolution at the species-level. More in detail, we conclude that:

I) Among the section *Beta*, the plastome sequences of *B. patula*, *B. vulgaris* subsp. *vulgaris* and *B. vulgaris* subsp. *maritima* are highly similar as indicated by the slightly lower bootstrap values (96/100) for these three beets. As this section harbors wild beets in relatively close geographical proximity across the coastal Mediterranean area, the detected similarity can be explained by (natural) crossing and gene flow due to close geographical proximity or accidental cross-pollination during cultivation as wild beet and cultivated beet groups are easily cross-compatible (3). Further, *B. vulgaris* subsp. *vulgaris* and some *B. vulgaris* subsp. *maritima* accessions are even phenotypically highly similar (41). The phylogenetic relationships among species can also be influenced by the geographical distribution, mating systems and polyploidization (10). Allogamy and self-incompatibility are characteristics of *B. vulgaris* subsp. *maritima*, whereas *B. macrocarpa* and *B. vulgaris* subsp. *adanensis* are self-compatible leading to lower divergence and higher homozygosity. Cross-compatibility can lead to hybridization by facilitating gene flow between individual species (40), especially between *B. patula* and *B. vulgaris* subsp. *maritima*, which may explain the lower bootstrap support and the more unclear relations in the phylogenetic tree presented here. Although previous studies found low divergence between *B. vulgaris* subspecies (10), *B. vulgaris* subsp. *adanensis* seems to be clearly separated from the other subspecies in our analysis, possibly explained by the geographical distance to the other investigated samples.

II) We confidently assigned specific species, including *B. nana*, to the section *Corollinae*: *B. corolliflora, B. intermedia, B. macrorhiza, B. lomatogona* and *B. nana* cluster together and form this section (also suggested by (7)). Here, we particularly focused on the *B.* section *Corollinae* by analysing the plastome sequences of eight accessions from five different species plus a biological replicate of the eponymous species *B. corolliflora. B. nana***,** which is endemic to Greece (1,7,42), was previously considered a separate *B.* section *Nanae***.** Our results, however, combined with multiple other studies, clearly show that *B. nana* falls within the *B.* section *Corollinae*, which is distinct from *B.* section *Beta* (7,9,13). In addition to our plastome-based phylogenetic analysis, further genomic evidence points to high genomic similarities between *B. nana* and other *Corollinae***:** For example, these species are marked by similar repeat accumulation profiles as shown for many individual transposable element types (43–46). Regarding plant characteristics, frost tolerance and seed hardiness are useful traits in the section *Corollinae***,** including *B. nana***,** but do not occur in any species of

the section *Beta* (47). Thus, frost tolerance is specific to the *Corollinae* when compared to the *Beta* and *Patellifolia* species. These points lead to the classification of *B. nana* as a member of the *Corollinae*. Considering the highly variable polyploid/hybrid status complexes within the *Corollinae*, our plant set encompassed three diploids (*B. macrorhiza*, *B. lomatogona*, and *B. nana*), a tetraploid (*B. corolliflora*), as well as a pentaploid (*B. intermedia*). Although the hybrid status and parental contributions of the polyploids remain unresolved (1,40), we present convincing evidence that *B. intermedia* and *B. corolliflora* are closely related. Thus, our plastome sequence analysis brings new evidence supporting the hypothesis that *B. corolliflora* and *B. intermedia* belong to a highly variable polyploid hybrid complex (summarised by (12); Figure 5). The investigation of the whole genome sequences of these polyploid species may help to resolve these parental contributions.

III) Among the *Patellifolia* members, *P. procumbens* and *P. webbiana* can be phylogenetically distinguished: *Patellifolia* was previously classified as *B.* section *Procumbentes* and there is still an ongoing taxonomic debate whether *P. patellaris*, *P. procumbens* and *P. webbiana* can be considered as separate species. However, due to molecular and morphological traits, *Patellifolia* are now mostly considered a separate genus which is divided into three distinct species (7,9,11). The relationships among the *Patellifolia* species could not be resolved in previous studies (1). In the phylogenetic tree presented here, *P. procumbens* and *P. webbiana* seem to be closely related (however still distinguished with high support) and clearly separated from the two *P. patellaris* accessions. The branch lengths distinguishing *B. patula* and the *B. vulgaris* subsp. *vulgaris*/*maritima*, which are both considered separate species, are highly similar (0.0001). The same branch length separates *P. procumbens* and *P. webbiana*. Therefore, our phylogenetic analysis indicates that the three *Patellifolia* species are distinguishable on a molecular level.

Comparing the results presented here with earlier studies, the previous investigation of Betoideae was substantially extended and refined. The phylogenetic relationships were resolved in more detail and not only based on the monophyletic groups. This is especially important for the species of the *B.* section *Corollinae* which were investigated in depth. Using the whole plastome sequences, including intergenic regions, it was possible to further resolve the phylogenetic relationships with higher bootstrap support due to the extraction of higher sequence variance and phylogenetic signal within the subfamily.

Conclusions

We provide 19 plastome assemblies for 18 different beet and wild beet accessions, which can also be re-used for future investigations of beets and other Caryophyllales species, and harnessed these to revisit systematic issues within the genera *Beta* and *Patellifolia*. This analysis advanced our understanding of the phylogenetic relationships of the subfamily Betoideae in four ways: I) Analysing sequences of intergenic regions of the whole plastome assemblies made it possible to reveal the phylogeny of closely related species with high reliability. Our phylogenetic tree shows a clear separation of the wild beet genera *Beta* and *Patellifolia*, as well as of the two sections *Beta* and *Corollinae*. II) *B. vulgaris* subsp. *adanensis* and *B. macrocarpa* can be clearly distinguished from *B. vulgaris* subsp. *vulgaris*, *B. vulgaris* subsp. *maritima* and *B. patula*. A clear split of *B. patula* from the two *B. vulgaris* subsp. (*B. vulgaris* subsp. *vulgaris* and *B. vulgaris* subsp. *maritima*) was not observed, likely due to the high sequence identity possibly explained by the close geographical proximity and the fact that these species are easily cross-compatible. III) All three *Patellifolia* species are clearly separated in our phylogenetic analysis, while *P. procumbens* and *P. webbiana* are more closely related to each other than to *P. patellaris*. These results, including the investigation of the branch lengths, point to a molecular separation within the *Patellifolia* species. IV) Finally, the taxonomic classification of *B. nana* as a member of the *Corollinae* was further supported.

Methods

Plant material, genomic DNA extraction, and DNA sequencing

Seeds of Betoideae species were obtained from KWS Saat SE, Einbeck, Germany (*B. vulgaris* subsp. *vulgaris* genotype KWS 2320) and from the Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany (all other accessions with accession numbers listed in Table 2 and Additional file 2). Apart from *B. vulgaris* subsp. *vulgaris*, 17 other *Beta* and *Patellifolia* accessions were analysed. The exact sampling location of the investigated accessions was extracted from the GBIS/I (Genebank information system; IPK) (48).

Table 2: Abbreviation, species name, accession number, genus, and section of the investigated accessions of the Betoideae subfamily.

Abbreviation	Species name	Accession number	Genus	Section
Bmar1	*Beta vulgaris* subsp. *maritima*	BETA 1101	*Beta*	*Beta*
Bada	*Beta vulgaris* subsp. *adanensis*	BETA 1233	*Beta*	*Beta*
Bmar2	*Beta vulgaris* subsp. *maritima*	BETA 2322	*Beta*	*Beta*
Bvul	*Beta vulgaris* subsp. *vulgaris*	KWS 2320	*Beta*	*Beta*
Bptu	*Beta patula*	BETA 548	*Beta*	*Beta*
Bmca	*Beta macrocarpa*	BETA 881	*Beta*	*Beta*
Bnan1	*Beta nana*	BETA 546	*Beta*	*Corollinae*, formerly *Nanae*
Bnan2	*Beta nana*	BETA 570	*Beta*	*Corollinae*, formerly *Nanae*
Bint1	*Beta intermedia*	BETA 431	*Beta*	*Corollinae*
Bint2	*Beta intermedia*	BETA 923	*Beta*	*Corollinae*
Bcor1	*Beta corolliflora*	BETA 408	*Beta*	*Corollinae*
Bcor2	*Beta corolliflora*	BETA 408	*Beta*	*Corollinae*
Bmrh1	*Beta macrorhiza*	BETA 830	*Beta*	*Corollinae*
Bmrh2	*Beta macrorhiza*	BETA 576	*Beta*	*Corollinae*
Blom	*Beta lomatogona*	BETA 674	*Beta*	*Corollinae*
Pweb	*Patellifolia webbiana*	BETA 526	*Patellifolia*	-
Ppro	*Patellifolia procumbens*	BETA 951	*Patellifolia*	-
Ppat1	*Patellifolia patellaris*	BETA 534	*Patellifolia*	-
Ppat2	*Patellifolia patellaris*	BETA 892	*Patellifolia*	-

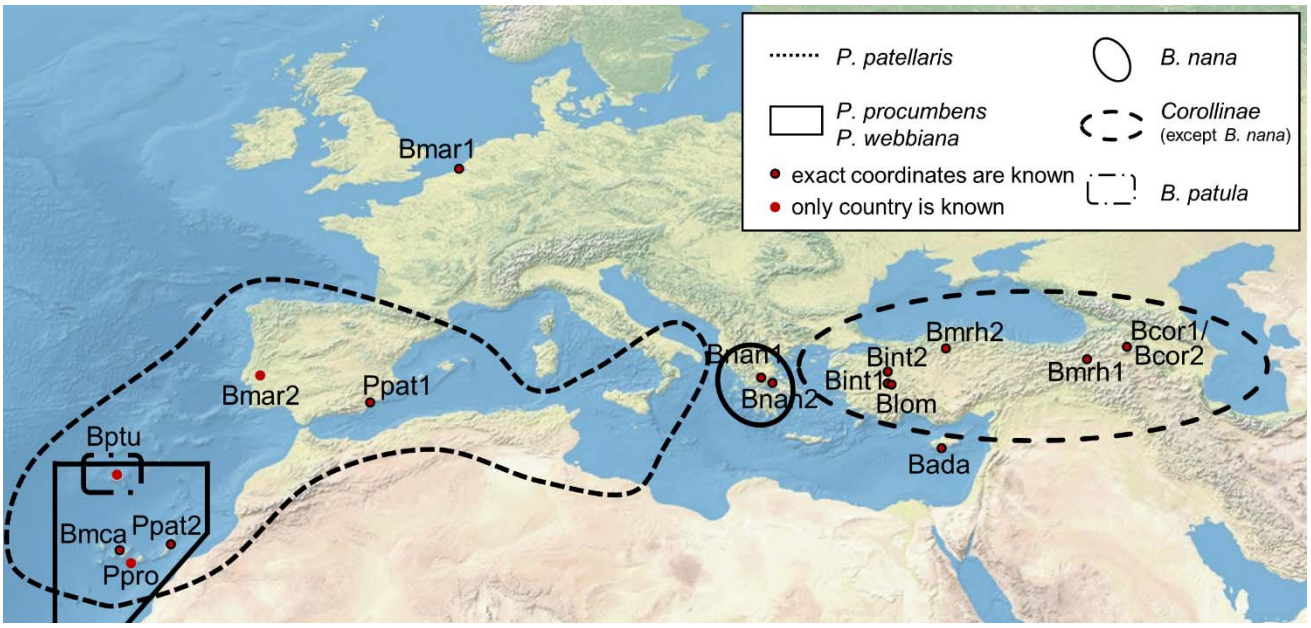


Figure 6: Geographic distribution of the *Beta* and *Patellifolia* species. The exact sampling locations of the investigated species are shown. The black lines represent the distribution area of the respective species and sections (see legend). The distribution areas of *B. vulgaris* subsp. *maritima*, *B. vulgaris* subsp. *adanensis*, and *B. macrocarpa* are not shown as these species occur along the whole coastline of Western Eurasia (1,9,10,49).

The plants were grown under long day conditions in a greenhouse and were obtained and grown in accordance with German legislation. Genomic DNA was isolated from young leaves using the NucleoSpin ® Plant II protocols from Macherey & Nagel. Each high-quality gDNA (200 ng) was fragmented by sonication using a Bioruptor (Fa. Diagenode) and subsequently used for library preparation with the TruSeq Nano DNA library preparation kit (Fa. Illumina). End repaired fragments were size selected by AmpureXp Beads (Fa. Beckmann-Coulther) to an average size of around 700 bp. After end repair, A-tailing and ligation of barcoded adapters, fragments were enriched by eight cycles of PCR. The final libraries were quantified using PicoGreen (Fa. Quant-iT) on a FLUOstar plate reader (Fa. BMG labtech) and quality checked by HS-Chip on a 2100 Bioanalyzer (Fa. Agilent Technologies). Before sequencing all libraries were pooled depending on the genome size and ploidy of each accession and sequenced 2 x 250 nt on a HiSeq1500 in rapid mode over two lanes using onboard cluster generation. Processing and demultiplexing of raw data was performed by bcl2fastq-v2.19.1 to generate FASTQ files for each accession.

Plastome assemblies and annotation

Trimmomatic (v0.39) (50) was applied to remove adapter sequences (ILLUMINACLIP:adapters.fa:2:30:10:2:keepBothReads) and to ensure high quality of the reads (SLIDINGWINDOW:4:15 MINLEN:50 TOPHRED33). FastQC (v0.11.9) (51) was used for quality checks. The trimmed reads were subjected to GetOrganelle (v1.7.0) (52) to generate plastome assemblies as suggested for Embryophyta plant plastome sequences (-R 15; -F embplant_pt). The SPAdes (53) kmer settings were set to -k 21, 45, 65, 85, or 105. The contig coverage information and other graph characteristics are used by GetOrganelle to construct the final assembly graphs, which were plotted and visually assessed using Bandage (v0.8.1) (54). The assemblies suggested a circular sequence, however, circular plastome molecules might only comprise a small proportion of all molecules in the cell, whereas other plastome molecules may occur in branched or linear configurations (55–57). The assemblies were submitted in the FASTA format, retaining the possibility to reuse the submitted assemblies as circular or linear sequences. The complex assembly graph of Bmar1 was not automatically resolved. Therefore, single contigs of Bmar1 were sorted manually based on the structure of the other assemblies to enable comparative analyses as described in the following section.

Structural annotation of all plastome assemblies was performed with GeSeq (v2.01) (58). The BLAT (59) search parameters ‘Annotate plastid trans-spliced rps12’ and ‘Ignore genes annotated as locus tags’ were used together with a ‘Protein search identity’ of 25 and a ‘rRNA, tRNA, DNA search identity’ of 85. For HMMER profile search ‘Embryophyta chloroplast (CDS+rRNA)’ was selected and ‘MPI-MP chloroplast references (Embryophyta CDS + rRNA)’ was chosen as reference. The resulting annotation files in the gff format were directly used for further analyses. To avoid confusion, we want to make aware of the fact that *psbN* and *pbf1* are two different names for the same gene.

Construction of phylogenetic trees

The workflow for the alignment-based phylogenetic analysis is available in Additional file S1G. The position of each gene was extracted from the GFF files obtained through GeSeq. Next, adjacent genes with conserved microsynteny across all investigated samples (including amaranth, quinoa, and spinach as outgroup) were identified and the interleaved

intergenic regions of these neighbouring genes were extracted. For overlapping genes, the extraction of an intergenic region was not possible.

Using the gene sequences and intergenic regions of all samples, gene/region specific alignments were performed using MAFFT (v7.299b) (60). High accuracy was ensured using the L-INS-I method. To align sequences with different orientations, the parameter '--adjustdirection' was used. The alignments were trimmed using trimAl (v1.4.rev22) (61). The gap threshold was set to '-gt 0.8', whereas the threshold for the minimum average similarity was set to '-st 0.001'. Then, the single alignments were concatenated and the resulting alignment matrix was inspected using SeaView (62). Manual adjustment was not necessary.

RAxML-NG (v1.0.0) (63) was used for ML analysis together with bootstrapping (Model: GTR+FC+G8m). The substitution matrix GTR (for DNA) was applied together with the model parameter 'G8' and 'F'. The parsimony-based randomised stepwise addition algorithm was selected for the starting tree (--tree pars{10}). The number of replicate trees for bootstrapping was set to 200. The resulting tree was visualised using FigTree (v1.4.4) (64).

Location-based clustering of the clades in the tree was performed manually based on the sampling locations (Additional file S2C).

To identify a reduced set of gene sequences and intergenic regions for the construction of a phylogenetic tree distinguishing all accessions, the sequences were iteratively added with increasing alignment identities until all species were separated by informative positions.

To investigate the region dependent phylogeny different additional data matrices were constructed for (I) all gene regions, (II) all intergenic regions, (III) complete coding sequences, (IV) first and second codon position and (V) third codon position only. Therefore, coding sequences for all 79 protein coding genes were extracted from the Genbank annotation files of our plastome assemblies. Start and stop codons were removed and extracted sequences were processed as described above for ML analysis.

To extract the 18s rRNA gene sequence from *B. corolliflora* (Bcor1) as representative of the *Corollinae*, SOAPdenovo2 assemblies were generated using the trimmed reads as input. SOAPdenovo2 (v2.04) (65) was tested with different kmer sizes ranging from 67-127 in

steps of 10. The resulting assembly with the highest N50 length was used for further investigations. The reference 18s rRNA gene sequence for *B. vulgaris* subsp. *vulgaris* was retrieved from the NCBI (GeneID=809573). The 18s rRNA gene sequence for Bcor1 was identified via BLAST and then extracted from the SOAPdenovo2 assembly, consecutively adding the following overlapping BLAST hit with the smallest e-value. Next, these extracted sequences were combined for a 18s gene sequence reconstruction. The assembled 18s rRNA gene sequence and the corresponding reference gene sequence were aligned via MAFFT and inspected using SeaView.

Mitogenome and plastome phylogeny based on kmer-derived phylogenetic splits

SANS serif (v2.1_04B) (66,67), a method based on colored de Bruijn graphs, was selected for the reconstruction of additional phylogenies using variable input data (mitochondrial reads, chloroplastic reads, and plastome assemblies). This method does not require prior assembly of the reads and is therefore especially suitable for the mitochondrial sequences which could not be fully assembled using GetOrganelle due to the relatively low available sequencing depth and also higher complexity of the mitogenome in comparison to the plastome.

Reads were assigned to the plastome or the mitogenome, respectively, after mapping with BWA-MEM (v0.7.13) (68) against the sugar beet reference genome sequence, including the respective chloroplast (KR230391.1) and mitochondrial sequence (BA000009.3). Therefore, ‘samtools view’, with the -b and -h options, was used after indexing the BAM file. The resulting BAM file was then further processed using ‘samtools collate’ and converted to the FASTQ format to extract the corresponding reads (chloroplast or mitochondria, respectively) using ‘samtools fastq’. After that, for the read-derived phylogenetic analyses, a colored de Bruijn graph was constructed using Bifrost (v1.0.5) (69) to filter kmers which only occur once in the dataset. This graph was then used as input for SANS serif. In addition, a phylogeny was reconstructed using the newly constructed plastome assemblies as direct input for SANS serif.

Different parameters were applied to test the robustness of the results. These arguments include different mean weight functions (-m; geom vs. geom2), the number of splits in the output list (-t; all vs. 10n) as well as various kmer sizes (11, 21, 31 and 61). The SANS serif output file was then converted to the nexus format (sans2nexus.py) and subsequently

visualised using Splitstree5 (70). To analyse the discrepancies between trees derived from different methods, SANS serif was used with the option ‘strict’ to generate an output file in the newick format which was then visualised using FigTree (64). Further, the SANS serif script ‘comp.py’ was used to calculate weighted (length of the edges/size of the splits is taken into account) precision and recall (combined in F1 scores) while using each tree as reference/ground truth in an ‘all vs. all’ comparison. In this use case, precision means ‘the total weight of all correctly predicted splits divided by the total weight of all predicted splits’. Further, weighted symmetric set distances and Robinsons-Foulds distances were calculated for each comparison. Details can be found in Additional file S1D. For this analysis, the trees constructed with different input data (cp_reads, mt_reads, and cp_assemblies) and the fixed parameters ‘-m geom2, -t 10n, -k 31’ were compared.

Investigation of alignment identities

The alignment identities for each plastome gene sequence or intergenic region were calculated to infer the phylogenetic information of all sequences. The events (SNV or InDel) were detected by iteration over each position in the sequence. The identity score (percent identity) was calculated by division of conserved positions (number of residues [position in alignment] with the same nucleotide in all accessions) by the number of residues in the alignment ('sequence length'). Alignment identities were calculated (i) for all accessions and (ii) for all accessions including outgroup reference plastome sequences (amaranth, quinoa, and spinach). Visualisation of the results was performed using matplotlib (71). Next, potential hotspots for SNVs and InDels in the plastome sequences were investigated.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Sequence reads have been submitted to the European Nucleotide Archive (ENA; Additional file S2D). The plastome assemblies and the corresponding annotations are available at ENA/GenBank (PRJEB45680). Additional information used in phylogenetic analyses are included in the supplementary files.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

Open Access funding enabled and organized by Projekt DEAL. KS is funded by Bielefeld University through the Graduate School DILS (Digital Infrastructure for the Life Sciences).

Authors' contributions

BP, BW, TH and DH designed the study. NS selected and cultivated the plants and performed DNA extraction. PV designed the layout for sequencing. BP and KS developed and implemented the bioinformatic methodology. KS analysed the data and prepared the figures and tables. KS, BP, NS and TH wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft (German Research Foundation) and the Open Access Publication Fund of Bielefeld University. We thank the CeBiTec Bioinformatic Resource Facility team for great technical support and Dr. Roland Wittler for great support with the SANS serif software.

References

1. Romeiras MM, Vieira A, Silva DN, Moura M, Santos-Guerra A, Batista D, et al. Evolutionary and Biogeographic Insights on the Macaronesian Beta-Patellifolia Species (Amaranthaceae) from a Time-Scaled Molecular Phylogeny. Robillard T, editor. PLoS ONE. 2016 Mar 31;11(3):e0152456.

2. Fischer HE. Origin of the ‘Weisse Schlesische Rübe’ (white Silesian beet) and resynthesis of sugar beet. Euphytica. 1989 Apr;41(1–2):75–80.

3. Panella L, Lewellen RT. Broadening the genetic base of sugar beet: introgression from wild relatives. Euphytica. 2007 Mar 7;154(3):383–400.

4. Biancardi E, Lewellen RT. History and Current Importance. In: Biancardi E, Panella LW, McGrath JM, editors. Beta maritima [Internet]. Cham: Springer International Publishing; 2020 [cited 2021 Jul 28]. p. 1–48. Available from: http://link.springer.com/10.1007/978-3-030-28748-1_1

5. Capistrano-Gossmann GG, Ries D, Holtgräwe D, Minoche A, Kraft T, Frerichmann SLM, et al. Crop wild relative populations of Beta vulgaris allow direct mapping of agronomically important genes. Nat Commun. 2017 Aug;8(1):15708.

6. Rodríguez del Río Á, Minoche AE, Zwickl NF, Friedrich A, Liedtke S, Schmidt T, et al. Genomes of the wild beets *Beta patula* and *Beta vulgaris* ssp. *maritima*. Plant J. 2019 Sep;99(6):1242–53.

7. Kadereit G, Hohmann S, Kadereit JW. A Synopsis of Chenopodiaceae Subfam. Betoideae and Notes on the Taxonomy of Beta. Willdenowia. 2006 Apr 20;Bd. 36, H. 1(Special Issue: Festschrift Werner Greuter):9–19.

8. Ford-Lloyd BV, Williams JT. A revision of Beta section Vulgares (Chenopodiaceae), with new light on the origin of cultivated beets. Botanical Journal of the Linnean Society. 1975 Sep;71(2):89–102.

9. Hohmann S, Kadereit JW, Kadereit G. Understanding Mediterranean□Californian disjunctions: molecular evidence from Chenopodiaceae□Betoideae. Taxon. 2006 Feb;55(1):67–78.

10. Touzet P, Villain S, Buret L, Martin H, Holl A-C, Poux C, et al. Chloroplastic and nuclear diversity of wild beets at a large geographical scale: Insights into the evolutionary history of the *Beta* section. Ecol Evol. 2018 Mar;8(5):2890–900.

11. Frese L, Nachtigall M, Iriondo JM, Rubio Teso ML, Duarte MC, Pinheiro de Carvalho MÂA. Genetic diversity and differentiation in Patellifolia (Amaranthaceae) in the Macaronesian archipelagos and the Iberian Peninsula and implications for genetic conservation programmes. Genet Resour Crop Evol. 2019 Jan;66(1):225–41.

12. Frese L, Ford-Lloyd B. Taxonomy, Phylogeny, and the Genepool. In: Biancardi E, Panella LW, McGrath JM, editors. Beta maritima [Internet]. Cham: Springer International Publishing; 2020 [cited 2021 Jul 28]. p. 121–51. Available from: http://link.springer.com/10.1007/978-3-030-28748-1_6

13. Shen Y, Ford-lloyd BV, Newbury HJ. Genetic relationships within the genus Beta determined using both PCR-based marker and DNA sequencing techniques. Heredity. 1998 May;80(5):624–32.

14. Gitzendanner MA, Soltis PS, Yi T-S, Li D-Z, Soltis DE. Plastome Phylogenetics: 30 Years of Inferences Into Plant Evolution. In: Advances in Botanical Research [Internet]. Elsevier; 2018

[cited 2021 Jul 28]. p. 293–313. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065229617300885

15. Liu B-B, Ma Z-Y, Ren C, Hodel RGJ, Sun M, Liu X-Q, et al. Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: A case study in Vitaceae [Internet]. Evolutionary Biology; 2021 Feb [cited 2021 Jul 28]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.02.25.432805

16. Palmer JD, Zamir D. Chloroplast DNA evolution and phylogenetic relationships in Lycopersicon. Proceedings of the National Academy of Sciences. 1982 Aug 1;79(16):5006–10.

17. Givnish TJ, Spalink D, Ames M, Lyon SP, Hunter SJ, Zuluaga A, et al. Orchid phylogenomics and multiple drivers of their extraordinary diversification. Proc R Soc B. 2015 Sep 7;282(1814):20151553.

18. Orton LM, Burke SV, Duvall MR. Plastome phylogenomics and characterization of rare genomic changes as taxonomic markers in plastome groups 1 and 2 Poeae (Pooideae; Poaceae). PeerJ. 2019 Jun 3;7:e6959.

19. Guo X, Liu J, Hao G, Zhang L, Mao K, Wang X, et al. Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics. 2017 Dec;18(1):176.

20. Singh BP, Kumar A, Kaur H, Singh H, Nagpal AK. CpGDB□: A Comprehensive Database of Chloroplast Genomes. Bioinformation. 2020 Feb 29;16(2):171–5.

21. Wang M, Wang X, Sun J, Wang Y, Ge Y, Dong W, et al. Phylogenomic and evolutionary dynamics of inverted repeats across Angelica plastomes. BMC Plant Biol. 2021 Dec;21(1):26.

22. Zurawski G, Clegg M. Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. Annual review of plant physiology. 1987;38:391–418.

23. Sugiura M. The chloroplast genome. Plant Molecular Biology (Netherlands). 1992;

24. Wang W, Lanfear R. Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants. Gaut B, editor. Genome Biology and Evolution. 2019 Nov 21;evz256.

25. Chen Y, Yang Z. Characterization of the complete plastome of *Dysphania botrys* , a candidate plant for cancer treatment. Mitochondrial DNA Part B. 2018 Jul 3;3(2):1214–5.

26. Yao G, Jin J-J, Li H-T, Yang J-B, Mandala VS, Croley M, et al. Plastid phylogenomic insights into the evolution of Caryophyllales. Molecular Phylogenetics and Evolution. 2019 May;134:74–86.

27. Stadermann KB, Weisshaar B, Holtgräwe D. SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome. BMC Bioinformatics. 2015 Dec;16(1):295.

28. Williams LE, Wernegreen JJ. Sequence Context of Indel Mutations and Their Effect on Protein Evolution in a Bacterial Endosymbiont. Genome Biology and Evolution. 2013 Mar;5(3):599–605.

29. Gomes Pacheco T, Morais da Silva G, de Santana Lopes A, de Oliveira JD, Rogalski JM, Balsanelli E, et al. Phylogenetic and evolutionary features of the plastome of Tropaeolum pentaphyllum Lam. (Tropaeolaceae). Planta. 2020 Aug;252(2):17.

30. Serna-Sánchez MA, Pérez-Escobar OA, Bogarín D, Torres-Jimenez MF, Alvarez-Yela AC, Arcila-Galvis JE, et al. Plastid phylogenomics resolves ambiguous relationships within the orchid family and provides a solid timeframe for biogeography and macroevolution. Sci Rep. 2021 Dec;11(1):6858.

31. Wang J-H, Moore MJ, Wang H, Zhu Z-X, Wang H-F. Plastome evolution and phylogenetic relationships among Malvaceae subfamilies. Gene. 2021 Jan;765:145103.

32. de Santana Lopes A, Pacheco TG, Santos KG dos, Vieira L do N, Guerra MP, Nodari RO, et al. The Linum usitatissimum L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. Plant Cell Rep. 2018 Feb;37(2):307–28.

33. Qiu T, Cui S. Evolutionary analysis for Phragmites ecotypes based on full-length plastomes. Aquatic Botany. 2021 Mar;170:103349.

34. Igea J, Juste J, Castresana J. Novel intron markers to study the phylogeny of closely related mammalian species. BMC Evol Biol. 2010;10(1):369.

35. Palmer JD, Herbon LA. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. Journal of Molecular evolution. 1988;28(1):87–97.

36. Heckenhauer J, Paun O, Chase MW, Ashton PS, Kamariah AS, Samuel R. Molecular phylogenomics of the tribe Shoreeae (Dipterocarpaceae) using whole plastid genomes. Annals of Botany. 2019 May 20;123(5):857–65.

37. Olmstead RG, Bedoya AM. Whole genomes: the holy grail. A commentary on: ‘Molecular phylogenomics of the tribe Shoreeae (Dipterocarpaceae) using whole plastidgenomes’. Annals of Botany. 2019 May 20;123(5):iv–v.

38. Kim Y-K, Jo S, Cheon S-H, Joo M-J, Hong J-R, Kwak M, et al. Plastome Evolution and Phylogeny of Orchidaceae, With 24 New Sequences. Front Plant Sci. 2020 Feb 21;11:22.

39. Walker JF, Walker-Hale N, Vargas OM, Larson DA, Stull GW. Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ. 2019 Sep 24;7:e7747.

40. Coons GH. The wild species of Beta. Proc Am Soc Sugar Beet Technol. 1954;8(2).

41. Biancardi E, de Biaggi M. Morphology. In: Biancardi E, Panella LW, McGrath JM, editors. Beta maritima [Internet]. Cham: Springer International Publishing; 2020 [cited 2021 Jul 29]. p. 61–86. Available from: http://link.springer.com/10.1007/978-3-030-28748-1_3

42. Frese L, de Carvalho MAP, Duarte C. Crop case study Beta L.(including Patellifolia AJ Scott et al.). AEGRO project. Julius Kühn-Institut, Bundesforschungsinstitut für Kulturpflanzen, Institut für Züchtungsforschung an landwirtschaftlichen Kulturen; 2011.

43. Gao D, Schmidt T, Jung C. Molecular characterization and chromosomal distribution of species-specific repetitive DNA sequences from *Beta corolliflora* , a wild relative of sugar beet. Genome. 2000 Dec 1;43(6):1073–80.

44. Heitkam T, Holtgräwe D, Dohm JC, Minoche AE, Himmelbauer H, Weisshaar B, et al. Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades. Plant J. 2014 Aug;79(3):385–97.

45. Maiwald S, Weber B, Seibt KM, Schmidt T, Heitkam T. The Cassandra retrotransposon landscape in sugar beet (*Beta vulgaris*) and related Amaranthaceae: recombination and re-shuffling lead to a high structural variability. Annals of Botany. 2021 Jan 1;127(1):91–109.

46. Weber B, Wenke T, Frömmel U, Schmidt T, Heitkam T. The Ty1-copia families SALIRE and Cotzilla populating the Beta vulgaris genome show remarkable differences in abundance, chromosomal distribution, and age. Chromosome Res. 2010 Feb;18(2):247–63.

47. Panella LW, Stevanato P, Pavli O, Skaracis G. Source of Useful Traits. In: Biancardi E, Panella LW, McGrath JM, editors. Beta maritima [Internet]. Cham: Springer International Publishing; 2020 [cited 2021 Jul 29]. p. 167–218. Available from: http://link.springer.com/10.1007/978-3-030-28748-1_8

48. Oppermann M, Weise S, Dittmann C, Knüpffer H. GBIS: the information system of the German Genebank. Database [Internet]. 2015 Jan 1 [cited 2021 Jul 28];2015. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bav021/2433153

49. Castro S, Romeiras MM, Castro M, Duarte MC, Loureiro J. Hidden diversity in wild Beta taxa from Portugal: Insights from genome size and ploidy level estimations using flow cytometry. Plant Science. 2013 Jun;207:72–8.

50. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

51. Andrews S. FastQC: a quality control tool for high throughput sequence data. [Internet]. 2020. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

52. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020 Dec;21(1):241.

53. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012 May;19(5):455–77.

54. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies: Fig. 1. Bioinformatics. 2015 Oct 15;31(20):3350–2.

55. Oldenburg DJ, Bendich AJ. Most Chloroplast DNA of Maize Seedlings in Linear Molecules with Defined Ends and Branched Forms. Journal of Molecular Biology. 2004 Jan;335(4):953–70.

56. Oldenburg DJ, Bendich AJ. DNA maintenance in plastids and mitochondria of plants. Front Plant Sci [Internet]. 2015 Oct 29 [cited 2021 Jul 28];6. Available from: http://journal.frontiersin.org/Article/10.3389/fpls.2015.00883/abstract

57. Shaver JM, Oldenburg DJ, Bendich AJ. The Structure of Chloroplast DNA Molecules and the Effects of Light on the Amount of Chloroplast DNA during Development in *Medicago truncatula*. Plant Physiology. 2008 Mar 3;146(3):1064–74.

58. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Research. 2017 Jul 3;45(W1):W6–11.

59. Kent WJ. BLAT---The BLAST-Like Alignment Tool. Genome Research. 2002 Mar 20;12(4):656–64.

60. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 2002 Jul 15;30(14):3059–66.

61. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972–3.

62. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and Evolution. 2010 Feb 1;27(2):221–4.

63. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Wren J, editor. Bioinformatics. 2019 Nov 1;35(21):4453–5.

64. Rambaut A. FigTree [Internet]. 2009. Available from: http://evomics.org/resources/software/molecular-evolution-software/figtree/

65. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. Erratum: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaSci. 2015 Dec;4(1):30.

66. Wittler R. Alignment- and reference-free phylogenomics with colored de Bruijn graphs. Algorithms Mol Biol. 2020 Dec;15(1):4.

67. Rempel A, Wittler R. SANS serif: alignment-free, whole-genome-based phylogenetic reconstruction. Schwartz R, editor. Bioinformatics. 2021 Jun 16;btab444.

68. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio] [Internet]. 2013 May 26 [cited 2021 Jul 28]; Available from: http://arxiv.org/abs/1303.3997

69. Holley G, Melsted P. Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. Genome Biol. 2020 Dec;21(1):249.

70. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution. 2006 Feb 1;23(2):254–67.

71. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9(3):90–5.

Supplementary Material

Additional file S1A: Geographic distribution of the Betoideae species as described in the literature.

Additional file S1B: Distribution of coverage values and assembly length (in bp) for each region and the total assemblies.

Additional file S1C: Circular and linear plots of selected plastome assembly sequences.

Additional file S1D: Distance metrics for the comparison of splitstree results.

Additional file S1E: Reduced phylogenetic tree based on 53 gene and intergenic regions.

Additional file S1F: Phylogenetic trees based on different sequence matrices.

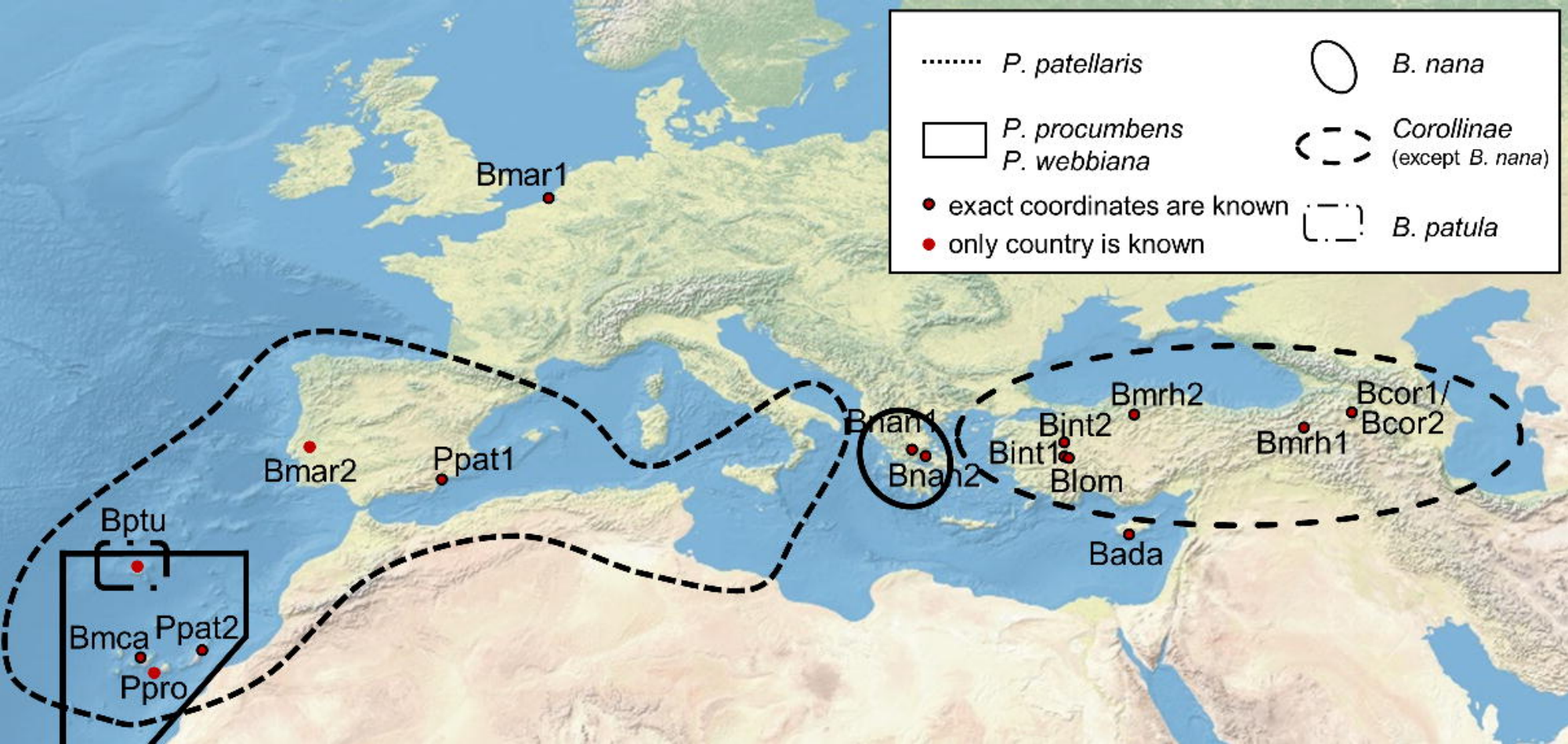
Additional file S1G: Workflow for the construction of phylogenetic trees.

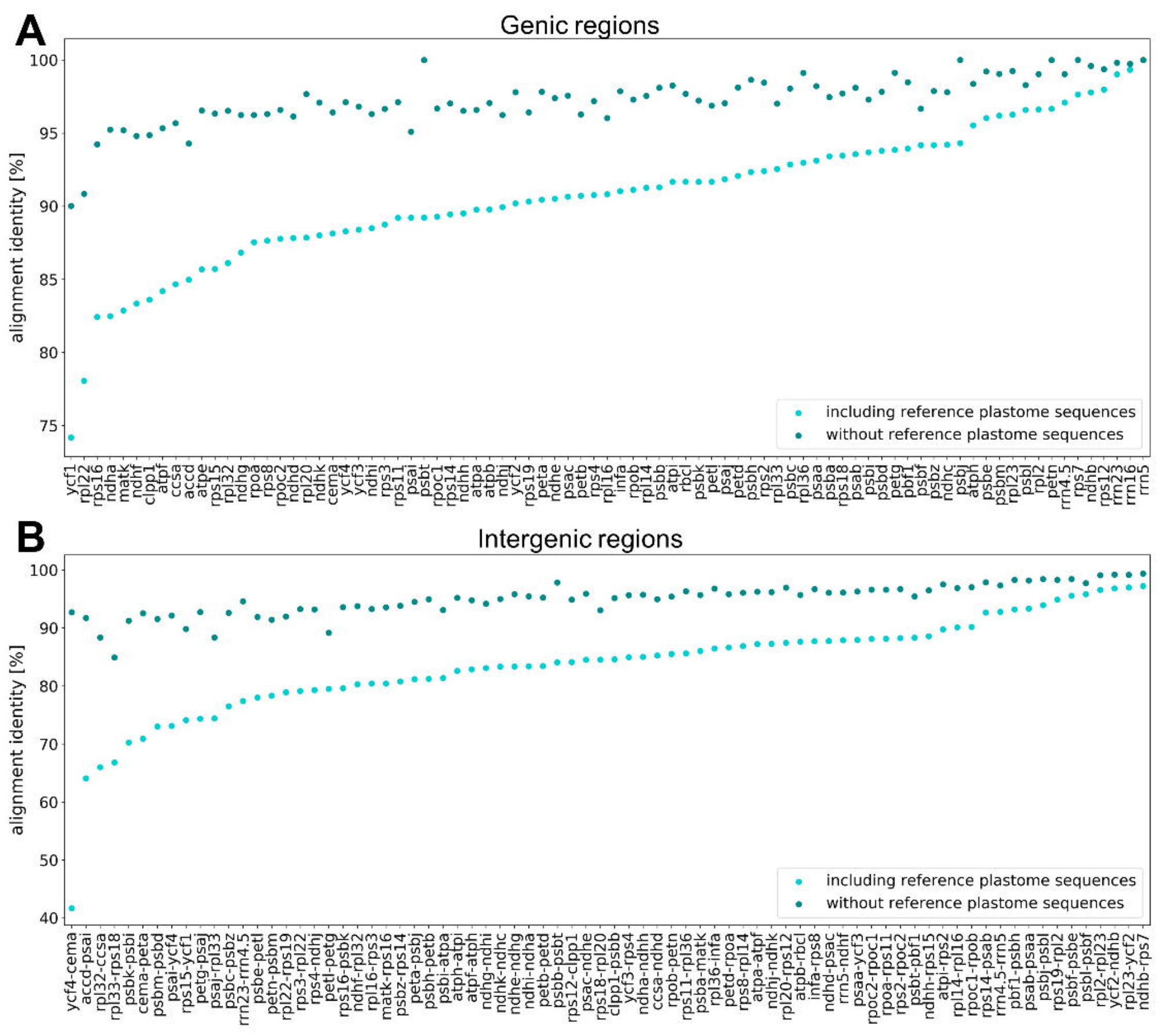
Additional file S2A: Sequence identities [%] of all investigated plastome gene sequences and intergenic regions including outgroup reference sequences.

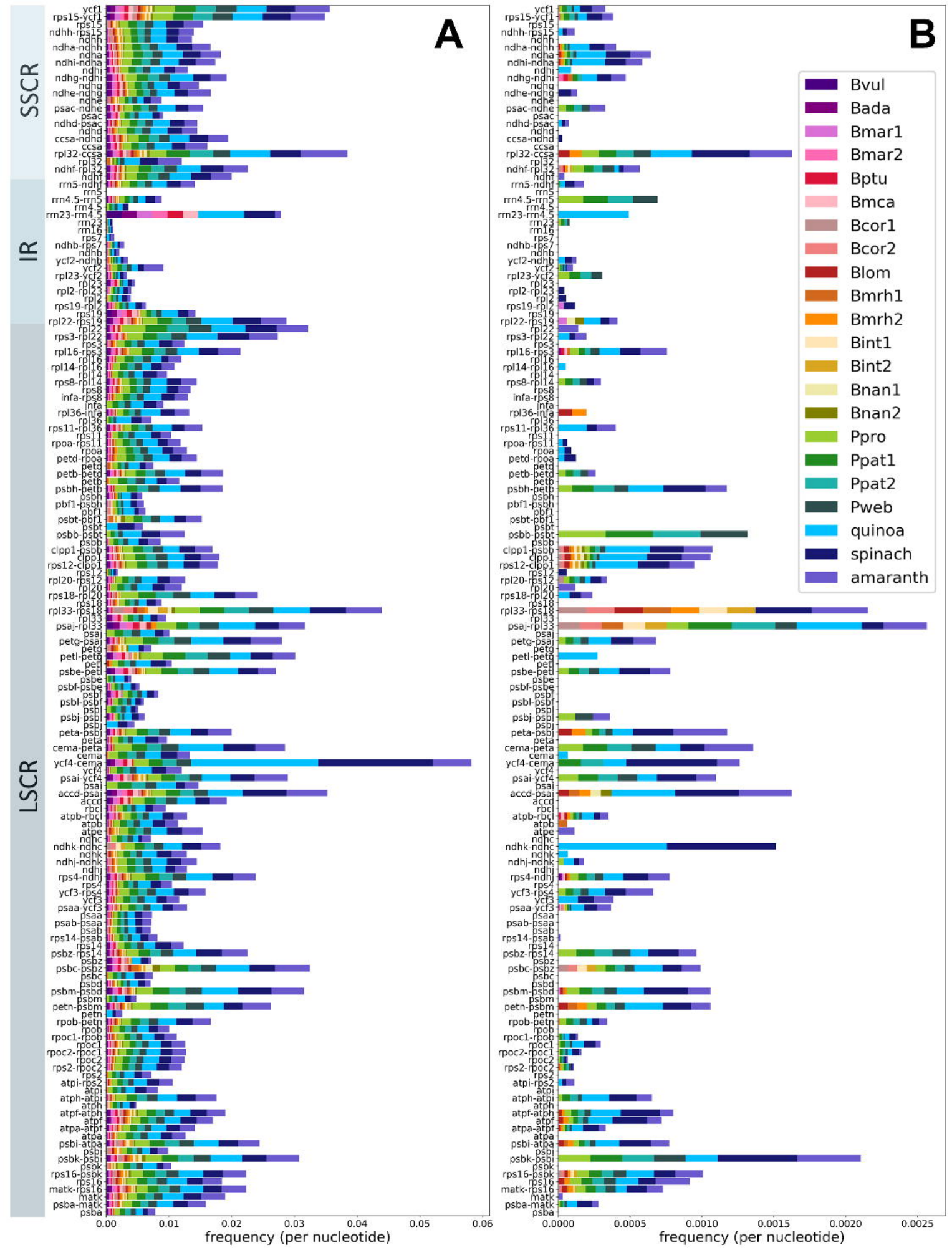
Additional file S2B: Sequence identities [%] of all investigated plastome gene sequences and intergenic regions excluding outgroup reference sequences.

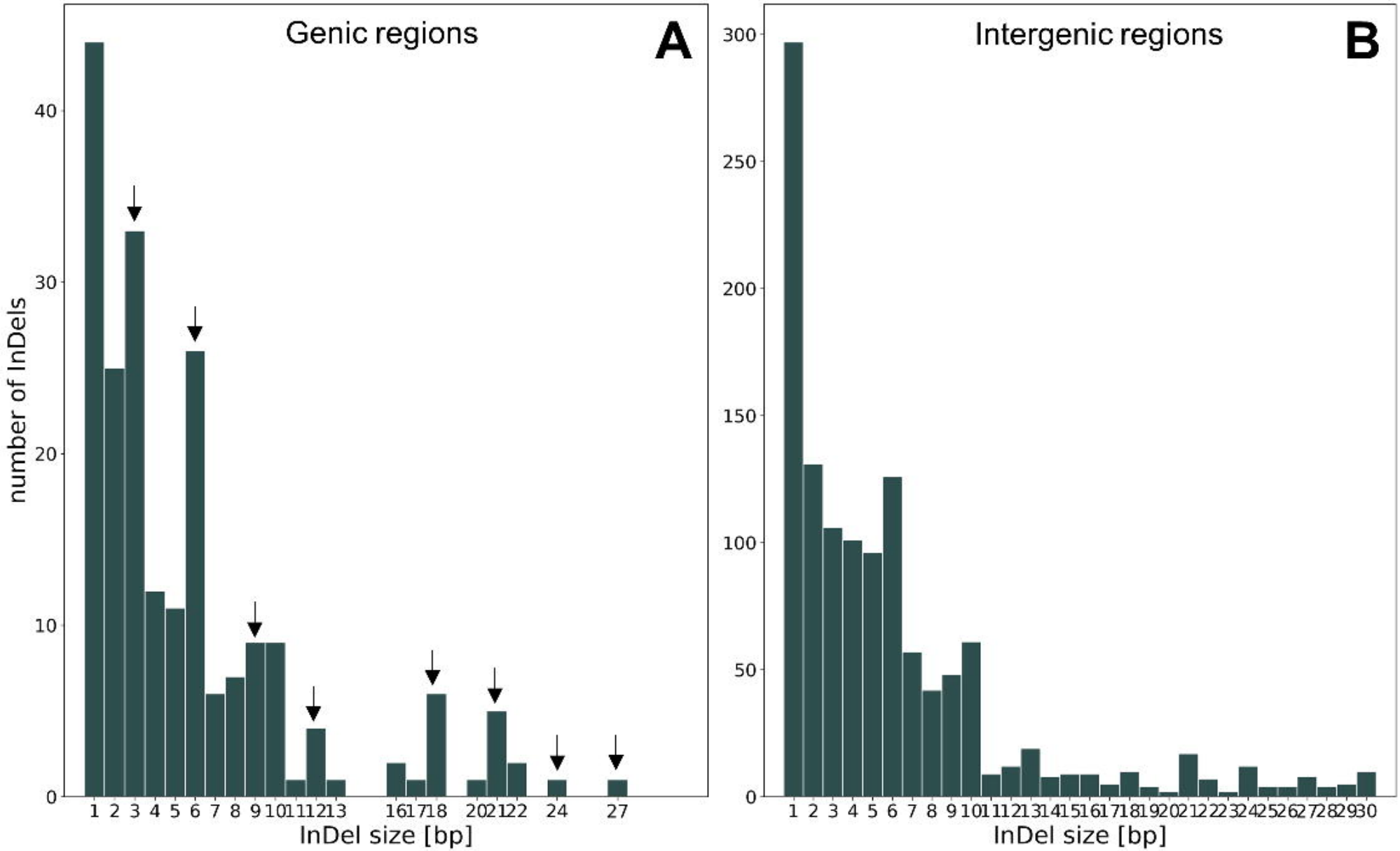
Additional file S2C: Accession IDs, taxonomy, read and assembly statistics and geographic location of the investigated accessions of the Betoideae subfamily.

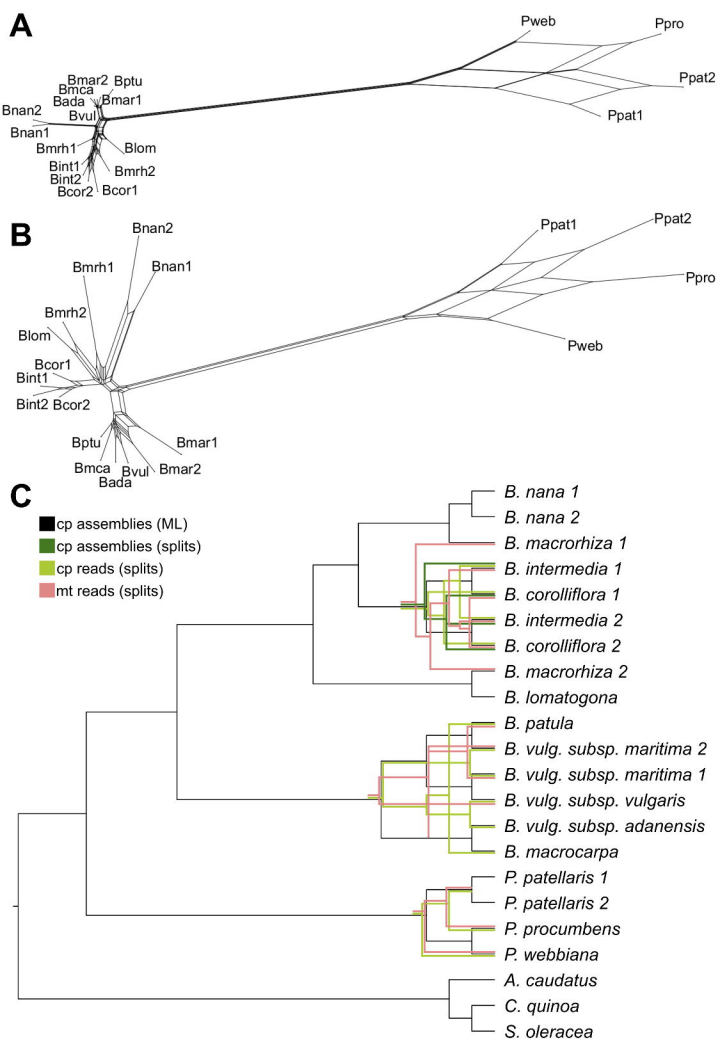
Additional file S2D: SRA-IDs of the processed read datasets.

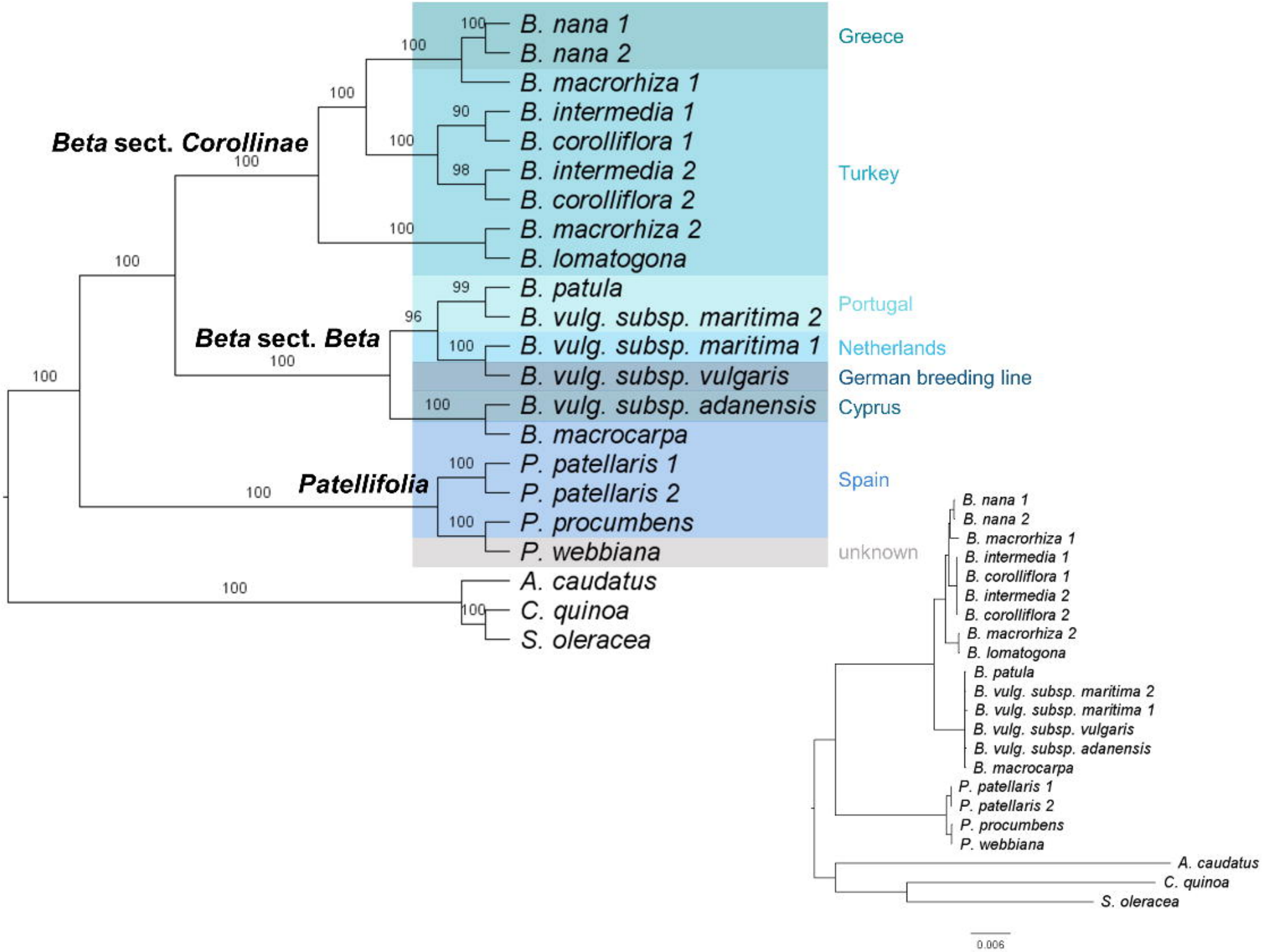

P. patellaris
B. nana
P. procumbens
P. webbiana
Corollinae
(except B. nana)
exact coordinates are known
only country is known
B. patula
Bmar1
Bmar2
Ppat1
Bnan1
Bnan2
Bint2
Bint1
Blom
Bmrh2
Bmrh1
Bcor1/
Bcor2
Bada
Bptu
Bmca
Ppat2
Ppro


A
Genic regions
alignment identity [%]
100
95
90
85
80
75
including reference plastome sequences
without reference plastome sequences
ycf1 rpl22 rps16 ndha matk ndhf clpp1 atpf ccsa accd atpe rps15 rpl32 ndhg rpoa rps8 rpoc2 ndhd rpl20 ndhk cema ycf4 ycf3 ndhi rps3 rps11 psai psbt rpoc1 rps14 ndhh atpa atpb ndhj ycf2 rps19 peta ndhe psac petb rps4 rpl16 infa rpob rpl14 psbb atpi rbcl psbk petl psaj petd psbh rps2 rpl33 psbc rpl36 psaa psba rps18 psab psbi psbd petg pbf1 psbf psbz ndhc psbj atph psbe psbm rpl23 psbl rpl2 petn rrn4.5 rps7 ndhb rps12 rrn23 rrn16 rrn5
B
Intergenic regions
alignment identity [%]
100
90
80
70
60
50
40
including reference plastome sequences
without reference plastome sequences
ycf4-cema accd-psai rpl32-ccsa rpl33-rps18 psbk-psbi cema-peta psbm-psbd psai-ycf4 rps15-ycf1 petg-psaj psaj-rpl33 psbc-psbz rrn23-rrn4.5 psbe-petl petn-psbm rpl22-rps19 rps3-rpl22 rps4-ndhj petl-petg rps16-psbk ndhf-rpl32 rpl16-rps3 matk-rps16 psbz-rps14 peta-psbj psbh-petb psbi-atpa atph-atpi atpf-atph ndhg-ndhi ndhk-ndhc ndhe-ndhg ndhi-ndha petb-petd psbb-psbt rps12-clpp1 psac-ndhe rps18-rpl20 clpp1-psbb ycf3-rps4 ndha-ndhh ccsa-ndhd rpob-petn rps11-rpl36 psba-matk rpl36-infa petd-rpoa rps8-rpl14 atpa-atpf ndhj-ndhk rpl20-rps12 atpb-rbcl infa-rps8 ndhd-psac rrn5-ndhf psaa-ycf3 rpoc2-rpoc1 rpoa-rps11 rps2-rpoc2 psbt-pbf1 ndhh-rps15 atpi-rps2 rpl14-rpl16 rpoc1-rpob rps14-psab rrn4.5-rrn5 pbf1-psbh psab-psaa psbj-psbl rps19-rpl2 psbf-psbe psbl-psbf rpl2-rpl23 ycf2-ndhb rpl23-ycf2 ndhb-rps7


A
B
SSCR
IR
LSCR
Bvul
Bada
Bmar1
Bmar2
Bptu
Bmca
Bcor1
Bcor2
Blom
Bmrh1
Bmrh2
Bint1
Bint2
Bnan1
Bnan2
Ppro
Ppat1
Ppat2
Pweb
quinoa
spinach
amaranth
0.00
0.01
0.02
0.03
0.04
0.05
0.06
frequency (per nucleotide)
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
frequency (per nucleotide)


Genic regions
A
number of InDels
0
10
20
30
40
1 2 3 4 5 6 7 8 9 10 11 12 13 16 17 18 20 21 22 24 27
InDel size [bp]
Intergenic regions
B
0
50
100
150
200
250
300
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
InDel size [bp]


A
Pweb
Ppro
Ppat2
Ppat1
Bmar2
Bptu
Bmca
Bada
Bmar1
Bnan2
Bvul
Bnan1
Bmrh1
Blom
Bint1
Bint2
Bmrh2
Bcor2
Bcor1
B
Bnan2
Ppat1
Ppat2
Bmrh1
Bnan1
Ppro
Bmrh2
Blom
Pweb
Bcor1
Bint1
Bint2
Bcor2
Bmar1
Bptu
Bmca
Bvul
Bmar2
Bada
C
cp assemblies (ML)
cp assemblies (splits)
cp reads (splits)
mt reads (splits)
B. nana 1
B. nana 2
B. macrorhiza 1
B. intermedia 1
B. corolliflora 1
B. intermedia 2
B. corolliflora 2
B. macrorhiza 2
B. lomatogona
B. patula
B. vulg. subsp. maritima 2
B. vulg. subsp. maritima 1
B. vulg. subsp. vulgaris
B. vulg. subsp. adanensis
B. macrocarpa
P. patellaris 1
P. patellaris 2
P. procumbens
P. webbiana
A. caudatus
C. quinoa
S. oleracea


Beta sect. Corollinae
Beta sect. Beta
Patellifolia
B. nana 1
B. nana 2
B. macrorhiza 1
B. intermedia 1
B. corolliflora 1
B. intermedia 2
B. corolliflora 2
B. macrorhiza 2
B. lomatogona
B. patula
B. vulg. subsp. maritima 2
B. vulg. subsp. maritima 1
B. vulg. subsp. vulgaris
B. vulg. subsp. adanensis
B. macrocarpa
P. patellaris 1
P. patellaris 2
P. procumbens
P. webbiana
A. caudatus
C. quinoa
S. oleracea
Greece
Turkey
Portugal
Netherlands
German breeding line
Cyprus
Spain
unknown
0.006