Variability among the Most Rapidly Evolving Plastid Genomic Regions is Lineage-Specific: Implications of Pairwise Genome Comparisons in Pyrus (Rosaceae) and Other Angiosperms for Marker Choice

Nadja Korotkova¹,²,³, Lars Nauheimer¹,², Hasmik Ter-Voskanyan³,⁴, Martin Allgaier⁵, Thomas Borsch¹,²,³

¹ Institut für Biologie/Botanik, Systematische Botanik und Pflanzengeographie, Freie Universität Berlin, Berlin, Germany, ² Dahlem Centre of Plant Sciences (DCPS), Berlin, Germany, ³ Botanischer Garten und Botanisches Museum Berlin-Dahlem, Berlin, Germany, ⁴ Institute of Botany, National Academy of Sciences of Republic Armenia, Yerevan, Armenia, ⁵ The Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Berlin, Germany

Abstract

Plastid genomes exhibit different levels of variability in their sequences, depending on the respective kinds of genomic regions. Genes are usually more conserved while noncoding introns and spacers evolve at a faster pace. While a set of about thirty maximum variable noncoding genomic regions has been suggested to provide universally promising phylogenetic markers throughout angiosperms, applications often require several regions to be sequenced for many individuals. Our project aims to illuminate evolutionary relationships and species-limits in the genus Pyrus (Rosaceae)—a typical case with very low genetic distances between taxa. In this study, we have sequenced the plastid genome of Pyrus spinosa and aligned it to the already available P. pyrifolia sequence. The overall p-distance of the two Pyrus genomes was 0.00145. The intergenic spacers between ndhC–tmnV, tmnR–atpA, ndhF–rpl32, psbM–trnD, and trnQ–rps16 were the most variable regions, also comprising the highest total numbers of substitutions, indels and inversions (potentially informative characters). Our comparative analysis of further plastid genome pairs with similar low p-distances from Oenothera (representing another rosid), Olea (asterids) and Cymbidium (monocots) showed in each case a different ranking of genomic regions in terms of variability and potentially informative characters. Only two intergenic spacers (ndhF–rpl32 and trnK–rps16) were consistently found among the 30 top-ranked regions. We have mapped the occurrence of substitutions and microstructural mutations in the four genome pairs. High AT content in specific sequence elements seems to foster frequent mutations. We conclude that the variability among the fastest evolving plastid genomic regions is lineage-specific and thus cannot be precisely predicted across angiosperms. The often lineage-specific occurrence of stem-loop elements in the sequences of introns and spacers also governs lineage-specific mutations. Sequencing whole plastid genomes to find markers for evolutionary analyses is therefore particularly useful when overall genetic distances are low.

Introduction

Clarifying species limits and reconstructing phylogenetic relationships in clades with recently diverged species is challenging. Levels of genetic divergence are often low while at the same time large numbers of samples need to be analysed. The same applies to analysing phylogeographic patterns, where many individuals from different populations need to be included. Due to the often complex modes of speciation in angiosperms, evidence from uniparentally inherited organellar genomes and the recombined nuclear genome is needed to unravel evolutionary histories [1–3]. This is also the case in the genus Pyrus where — like in many Rosaceae — polyploidy, hybridization, and reticulate evolution occur. Estimates of Pyrus diversity vary between 50 and 80 species [4,5] and 20 taxa alone have been described from the southern Caucasus [6,7]. Similarly, the numbers of accepted species differ between treatments as a consequence of poorly understood species limits. Pyrus is a typical case for evolutionary and taxonomic analyses of diverse species groups in flowering plants that require the inclusion of hundreds of individuals. Before entering into largescale sampling, we were interested to find the genomic regions with the best information potential for generating haplotype networks and inferring phylogenetic relationships. In this study, we focus on the plastid genome.
Alone the same line of argumentation, Shaw et al. [8,9] inspired to employ a broader spectrum of noncoding and rapidly evolving plastid markers in phylogenetic analyses of closely related species. Shaw et al. [8] sequenced a wide range of plastid markers for three species across angiosperms and later compared plastid genome pairs of three lineages of angiosperms (Atropa and Nicotiana for the asterids, Lotus and Medicago for the rosids, and Oryza and Saccharum for the monocots) [9]. Their studies resulted in a set of 32 regions that ranked highest in their number of potentially informative characters (defined as sum of substitutions, indels and inversions following [8] and abbreviated as “PICs”). This set was consequently suggested to generally contain the most variable and phylogenetically most informative genomic regions in angiosperm plastid genomes. However, the question remains how to best select four or five of the total top 32 regions, as many species-level evolutionary studies require.

Noncoding genomic regions such as introns and spacers often contain stem-loops and other specific structural elements that can be highly dynamic and are AT-rich. This results in a mosaic-like pattern of conserved and variable elements [10]. Considering that certain stem-loop elements within given introns and spacers are often unique to restricted lineages [11,12], lineage specificity in the overall variability of genomic regions is to be expected. In several recent comparative analyses of angiosperm plastid genomes [13,14] different genomic regions were depicted as the most variable. Nonetheless, these results need to be considered with care because some of the respective authors worked with pairs of hardly variable. Nonetheless, these results need to be considered with care. Nevertheless, these results need to be considered with care because some of the respective authors worked with pairs of hardly differentiated genomes while others had pairs of genomes with high p-distances. We expect that taxon-specific differences caused by certain sequence elements will be less prominent when more distant genomes are studied.

Next-generation sequencing techniques greatly facilitate the analysis of whole plastid genomes [15–17]. To date, phylogenomic studies of plastid genomes in land plants often just relied on concatenated sequences of the conserved genes, neglecting the information from the noncoding regions. In other cases, the authors included rather few taxa for which plastid genome sequences were automatically assembled from the respective 454 or Illumina runs, without completing parts of low coverage or areas with difficulties to obtain correct sequences. However, especially those might be informative at and below the species level (e.g., AT-rich stretches of DNA including microsatellites) [18–20]. On the other hand, there are recent studies which used completely annotated plastid genomes to detect infraspecific variability in species of Olea [21], Colocasia [22], or Phalaenopsis [23], to find genomic regions with the highest number of potentially informative characters in more distant genome pairs of angiosperm genera [9,24–26].

We have sequenced the plastid genome of Pyrus spinosa using 454 pyrosequencing in order to compare it with the published plastid genome sequence of P. pyrifolia [27]. In our Pyrus genome pair, the proportion of sites at which the two sequences are different (p-distances) is almost 10-fold lower than in the genome pairs studied by Shaw et al. [9]. For further comparison, we selected three fully annotated plastid genome pairs using the criterion of low p-distances (≤0.005) similar to Pyrus. Here we wanted to represent another rosid pair (Oenothera parviflora and O. argillicola; Onagraceae), an asterid pair (Olea europaea and O. woodiana; Oleaceae) and a monocot pair (Cymbidium tortisepalum and G. sinense (Orchidaceae)).

The goals of this study were (1) to find the most variable regions of the Pyrus plastid genome and to propose plastid markers for species-level evolutionary studies in Pyrus, (2) to assess the variability of plastid genome regions based on comparable genome-pairs with overall low p-distances (0.0005 to 0.005) in major lineages of angiosperms, (3) to clarify if there are universal or lineage-specific rankings of variability within the group of about 35 top variable genomic regions, and (4) to evaluate if there are lineage specific differences in molecular evolutionary patterns that could cause the variability of genomic regions.

**Material and Methods**

DNA extraction, 454 pyrosequencing, genome assembly and annotation

*Pyrus spinosa* was sampled from the living collection of the Botanical Garden Berlin-Dahlem (Acc. No. 24845810, IPEN-Nr. TR-0-B-2484581, origin: Turkey, Kastamonu, Pontic Mountains around Küre, leg.: Ern, Krone 7145, 9/1981, voucher at B). The leaf tissue was silica-dried and total genomic DNA was extracted using the NucleoSpin Plant II kit (Macherey Nagel) according to the manufacturer’s instructions. shotgun sequencing from total genomic DNA was performed on a Roche 454 GS-FLX Titanium sequencer (Roche Applied Science, Indianapolis, Indiana, USA). The 454 run (1/4 plate) resulted in 120,255 reads with an average of 400 bp after removing the adaptor sequences.

An initial mapping assembly with MIRA 4 [31] using *Pyrus pyrifolia* as reference resulted in 4191 reads mapped to a single contig with an average coverage of 13.44. However, reads with larger indels, not occurring in the reference, were not incorporated into the contigs what lead to an incorrect genome sequence. To remove the bias of the reference sequence, the reads were de novo assembled to contigs using the Roche GS *De Novo* Assembler (Newbler) v.2.6 which resulted in 836 large contigs (N50 = 829), and with Mira 4 [28], which resulted in 1125 large contigs (N50 = 1072, N90 = 538, N95 = 519). All these contigs were mapped on the *Pyrus pyrifolia* plastid genome (GenBank acc. no. NC015996; Terakami et al. [27]) using Geneious 7 to produce a consensus sequence. The combined method of mapping de novo contigs recovered nine indels (maximum length 71 bp), which were not found with mapping alone. Finally the second inverted repeat was manually inserted into the consensus sequence.

The positions of protein coding genes, tRNAs, tRNAs and the inverted repeats were annotated with the help of DOGMA [29] and Geneious 7. All coordinates of exons, reading frames and the positions of tRNAs were manually checked by aligning the respective genes (electrophoresis was done at Macrogen Europe, Amsterdam, the Netherlands). The 454 run (1/4 plate) resulted in 1125 large contigs (N50 = 829), and with Mira 4 [28], which resulted in 1125 large contigs (N50 = 1072, N90 = 538, N95 = 519). All these contigs were mapped on the *Pyrus pyrifolia* plastid genome (GenBank acc. no. NC015996; Terakami et al. [27]) using Geneious 7 to produce a consensus sequence. The combined method of mapping de novo contigs recovered nine indels (maximum length 71 bp), which were not found with mapping alone. Finally the second inverted repeat was manually inserted into the consensus sequence.

Verification by Sanger sequencing. Pyrosequencing is limited in that the exact number of nucleotides within longer homonucleotide stretches (polyAs or polyTs) cannot be reliably determined [16,31]. Our initial assembly contained several homonucleotide stretches and AT-rich sequence motifs. In our data, ambiguously called bases were frequent in homonucleotide stretches with more than six of the same nucleotides. To validate the sequence in such parts, we applied the Sanger method (electrophoresis was done at Macrogen Europe, Amsterdam, The Netherlands). Primers for amplification and sequencing were taken from the literature or designed in this study (see Table S1). Phoregrams were checked by eye for peaks and corresponding quality scores to ensure that the polyA/T stretch was correctly read. All Sanger sequencing reads were unambiguous with no overlapping peaks after the polyA/T stretches. The respective
reads were aligned with the previously assembled genome sequence in Genious 7 and the consensus sequence was corrected accordingly. The *Pyrus spinosa* plastid genome sequence is available in EMBL under accession HG737342.

**Pairwise genome comparisons and calculation of sequence divergence.** In addition to *Pyrus*, we took three other plastid genome pairs from published sources to represent closely related species, a further rosid genus, an asterid and a monocot genus. Genome sequences had to be complete and fully annotated. The aligned genome pairs had to show an overall distance of \( p \leq 0.005 \) (Table 1). All genome sequences were aligned in PhyDe using a motif alignment approach [32,33]. The pairwise alignments are provided as File S1, S2, S3, and S4.

Sequences of all introns and intergenic spacers larger than 100 bp were extracted from the alignments. The number of single nucleotide polymorphisms (SNPs) and indels for each sequence pair were counted with a script in R (v. 3.0.2). PICs were then calculated in intervals of increasing size (1–10, 20, 50, and 100 bp in each direction). A script was written in R v.3.0.2, which distinguishes the indels and regions around SNPs, calculates the AT contents, and displays their distributions.

Molecular evolution within genomic regions

In order to assess the role of the base composition in variable sequence parts, i.e., indels and nucleotides around SNPs, we calculated their AT contents and compared them with the overall AT content of the whole genomes (consensus of pairwise aligned genomes). Three groups of indels were distinguished: (1) length variable poly-n loci that consist of a single nucleotide that is repeated at least sevenfold, (2) simple sequence repeats (SSRs) that show one repetition of a motif of multiple nucleotides, inverted repeats, or inversions, and (3) indels that do not fall in the former categories.

Further, AT contents of nucleotides adjacent to SNPs were calculated in intervals of increasing size (1–10, 20, 50, and 100 bp in each direction). The lineage-specific occurrence of substitutions and microstructural mutations was examined in more detail on the example of group II introns (*atpF, rpl16*) that strongly deviated in variability among our four genome pairs. These introns possess a mosaic-like structure of conserved and variable sequence elements. The variable parts usually correspond to the structurally and functionally least constrained terminal stem-loops, which appear in the respective RNA secondary structure. We first annotated the domains of the *atpF* and *rpl16* introns by comparing our sequences with the consensus alignment of Michel et al. [37]. The RNA secondary structures of individual domains were then predicted using RNAstructure 5.6 (available at http://rna.urmc.rochester.edu/RNAstructure.html) using the algorithm of Mathews et al. [38]. The “fold as RNA” option was implemented to allow for U–G pairings.

Selecting genomic regions as markers for evolutionary studies in *Pyrus*. Our aim was not only to find the most variable plastid regions in *Pyrus* but also to select several regions to be best used in evolutionary studies of *Pyrus*. Thus, efficient

---

**Table 1.** GenBank accession numbers and references for the plastid genomes used in this study.

| Species                  | GenBank accession number | Reference                  |
|--------------------------|--------------------------|----------------------------|
| *Pyrus spinosa*          | HG737342                 | this study                 |
| *Pyrus pyrifolia*        | NC015996                 | Terakami et al. [27]       |
| *Cymbidium tortisepalum* | NC021431                 | Yang et al. [24]           |
| *Cymbidium sinense*      | NC021430                 | Yang et al. [24]           |
| *Oenothera parviflora*   | NC010362                 | Greiner et al. [66]        |
| *Oenothera argillicola*  | EU262887                 | Greiner et al. [67]        |
| *Olea woodiana*          | NC015608                 | Besnard et al. [68]        |
| *Olea europaea*          | NC015401                 | Besnard et al. [68]        |

**Table 2.** Sequence statistics for the four genome pairs compared.

| Genome pair                  | p-distance | Aligned length [bp]       | Length difference | SNPs | Indels |
|------------------------------|------------|---------------------------|-------------------|------|--------|
| *Pyrus spinosa*/*P. pyrifolia* | 0.00145    | 160607 bp                  | 227 bp            | 230  | 173    |
| *Olea europaea*/*O. woodiana* | 0.00294    | 156091 bp                  | 30 bp             | 458  | 112    |
| *Oenothera parviflora*/*O. argillicola* | 0.00122 | 165952 bp                  | 199 bp            | 173  |        |
| *Cymbidium tortisepalum*/*C. sinense* | 0.0008   | 155833 bp                  | 79 bp             | 124  | 62     |

PLOS ONE | www.plosone.org 3 November 2014 | Volume 9 | Issue 11 | e112998
amplification and sequencing strategies including primer binding sites, region size and the information content per primer read had to be considered in addition to a high rank in terms of variability. Furthermore, polyA/T stretches larger than seven nucleotides (microsatellites) had to be considered. Their presence usually require two primer reads for sequencing that start from both ends of the amplicon because slippage is likely to occur after the polyA/T stretch. Since a region >1000 bp usually requires two primers to sequence, one microsatellite was not considered a problem, while several microsatellites within the same region led to dismiss it. Considering that current technology generates reliable read lengths of 800–1000 bases, we selected fragments of 900–1300 bp in size—a size range that can be easily amplified and then sequenced with a maximum of two primers.
Results and Discussion

Size and structure of the *Pyrus* plastid genome

The plastid genome of *Pyrus spinosa* is 159,694 bp in length, and the inverted repeats (IRs) account for 26,396 bp. The large single-copy region (LSC) is 87,694 bp in length and the small single-copy region (SSC) 19,205 bp. The genome has a GC content of 36.6%. Gene content and order are identical to *Pyrus pyrifolia*, with 113 unique genes and 17 duplicates in the IR [30]. The extension of IRs is identical to *P. pyrifolia*, while a 137 bp gap in the LSC of *P. spinosa* directly adjacent to IRa leads to a different IR boundary. The p-distance between the two genomes is 0.00145 (Table 2). The consensus structure of the two *Pyrus* genomes and the variability between them is illustrated in Fig. 1. Most of the variation occurs in the noncoding parts, especially in intergenic spacers of the LSC region. The SSC is less variable and

---

Figure 2. Circular representation of plastid genome pair in *Cymbidium*. Shown are consensus sequences of compared species pairs of *Cymbidium tortisepalum* and *C. sinense* with their differing p-distances, numbers of SNPs and indels across the consensus. Radial grey highlights show the regions in focus of study with their names. Circular graphs from outside to inside: outermost circle with ticks for every 1,000 bp (small) and 10,000 bp (big) indicates part of genome, single copy regions in light grey and inverted repeats in dark grey; bands show locations of genes (blue), tRNAs (yellow) and rRNAs (red); the three outermost histograms display p-distances (blue), number of SNPs (green) and indels (orange) per spacer region; innermost graph shows number of SNPs (green histogram), indels (orange histogram), and AT content relative to the whole consensus (black line graph) of 500 bp long parts of the whole consensus. doi:10.1371/journal.pone.0112998.g002

---

*Pyrus* Genome Comparison
| Region       | Rank | Aligned length | PICs (SNPs/aligned length) | PICc (SNPs/aligned length) | Aligned length | PICs (SNPs/aligned length) | PICc (SNPs/aligned length) |
|--------------|------|----------------|---------------------------|---------------------------|----------------|---------------------------|---------------------------|
| psbB–psbT    | 1    | 184            | 6 (5/1)                   | 23.82                     | 194            | 6 (5/1)                   | 21.89                     |
| ndhF–rpl32   | 2    | 760            | 24 (12/12)                | 20.34                     | 20 (10/10)     | 17.70                     | 20.34                     |
| psbI–trnS    | 3    | 613            | 13 (12/1)                 | 12.15                     | 12 (10/2)      | 12.15                     | 12.15                     |
| ndhF–rpl32   | 4    | 909            | 20 (10/10)                | 10.89                     | 20 (10/10)     | 10.89                     | 10.89                     |
| psbC–trnS    | 5    | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 6    | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 7    | 1078           | 184 (12/12)               | 37.88                     | 184 (12/12)    | 37.88                     | 37.88                     |
| psbB–psbT    | 8    | 1089           | 18 (11/7)                 | 12.15                     | 18 (11/7)      | 12.15                     | 12.15                     |
| ndhF–rpl32   | 9    | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| psbC–trnS    | 10   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 11   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 12   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| psbC–trnS    | 13   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 14   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 15   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| psbC–trnS    | 16   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 17   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 18   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| psbC–trnS    | 19   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 20   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 21   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| psbC–trnS    | 22   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 23   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 24   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| psbC–trnS    | 25   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 26   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 27   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| psbC–trnS    | 28   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| trnR–atpA    | 29   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |
| ndhF–rpl32   | 30   | 134            | 2 (2/0)                   | 14.93                     | 2 (2/0)        | 14.93                     | 14.93                     |

The regions are sorted according to PIC-distances. P-values are calculated with Wilcoxon rank sum test.
almost no variation is found in the IRs. There are some genome parts with intergenic spacers alternating tRNA genes where variation appears to accumulate. This is especially the case in the region from \(\text{trnK}\) to \(\text{trnA}\) and from \(\text{rpsB}\) to \(\text{psbD}\) (Figs. 1, 2).

**Finding the most variable regions of the Pyrus plastid genome.** The five regions with the highest \(p\)-distances are the intergenic spacers \(\text{psbB–psbT}\), \(\text{psbI–trnS}\), \(\text{ndhC–trnV}\), \(\text{trnR–atpA}\), and \(\text{ndhF–rpl32}\). Taking the PICs as a basis, the five top-ranked regions are \(\text{ndhC–trnV}\), \(\text{trnR–atpA}\), \(\text{ndhF–rpl32}\), \(\text{psbM–trnD}\), and \(\text{trnQ–rps16}\) (Table 3, Fig. 1–4).

Comparing our results with the ranking of Shaw et al. [9] it appears that 17 of our 30 top-ranked regions in *Pyrus* are also among the 32 top-ranked in their study. However, their ranks are different. For example, in Shaw et al. [8], the \(\text{rpl32–trnL}\) spacer...
has the highest number of PICs whereas it is only at rank 8 in *Pyrus*. The *trnR–atpA* spacer, which has the second-highest number of PICs in *Pyrus*, was not at all reported. However, the ranking of Shaw et al. may not be that comparable because the authors “normalized” their PICs with the aim to reduce the influence of different evolutionary rates or genetic distances. They divided the number of PICs within a region from a certain taxonomic lineage by the total sum of PICs within the same lineage. Therefore, their results do not directly show lineage-specific differences in marker variability, although the absolute variability of a given genomic region is the only relevant fact in any analysis.

Low genetic distances in *Pyrus* have been pointed out in two earlier studies of *Pyrus* plastid genomes [27,39]. These studies

---

**Figure 4. Circular representation of plastid genome pairs in Olea.** Shown are consensus sequences of compared species pairs of *Olea europaea* and *O. woodiana* with their differing *p*-distances, numbers of SNPs and indels across the consensus. Radial grey highlights show the regions in focus of study with their names. Circular graphs from outside to inside: outermost circle with ticks for every 1,000 bp (small) and 10,000 bp (big) indicates part of genome, single copy regions in light grey and inverted repeats in dark grey; bands show locations of genes (blue), tRNAs (yellow) and rRNAs (red); the three outermost histograms display *p*-distances (blue), number of SNPs (green) and indels (orange) per spacer region; innermost graph shows number of SNPs (green histogram), indels (orange histogram), and AT content relative to the whole consensus (black line graph) of 500 bp long parts of the whole consensus.

doi:10.1371/journal.pone.0112998.g004
were motivated by the horticultural importance of Pyrus, and focused on Asian species and cultivars. Katayama and Uematsu [39] provided a physical map of the plastid genome of Pyrus ussuriensis var. hondoensis and ran an RFLP analysis on cpDNAs from 11 accessions of five Pyrus and two Prunus species. However, there were no sequence data to support their conclusions. Terakami et al. [27] aligned the three plastid genomes of Pyrus pyrifolia, Malus × domestica, and Prunus persica. The authors calculated the proportion of mutational events using the same formula as Shaw et al. [9] for 89 noncoding regions, and ranked the compared regions according to their variability comparing Pyrus with Malus and Prunus (ingroup and outgroup were not specifically defined). While the ndhC-trnV and trnR-accD spacers depict the highest sequence divergence in both, Terakami et al. and our work presented here, the overall rankings are strongly different. Terakami et al. found the spacers rpl33-psbA, psbI-trnS, and rpl14-rps16 from the third to fifth rank. In our Pyrus ranking, these spacers are at positions 22, 2, and 12 (based on p-distances) and 43, 22, and 41 (based on PICs), respectively. These differences may be explained by the much greater distance between the Pyrus and Malus plastid genomes than our two Pyrus genomes. The crown group of Pyrus diversified 27–33 mya while the crown group of Malus was inferred to have diversified 34–46 mya [40].

Various plastid regions have also been sequenced for a large number of samples in Pyrus. Katayama et al. [41] sequenced the rps16-trnQ and accD-psiA spacers and reconstructed a network based on 25 different haplotypes including 21 species of Pyrus and multiple individuals of P. pyrifolia and P. ussuriensis, respectively. The authors found both spacers to contain highly variable AT-rich mutational hotspots and concluded that these regions are “hypovariable”, while their remaining Pyrus sequences showed hardly any variation. The authors argued that their results confirmed their earlier hypothesis of strong sequence conservation in the plastid genomes of Pyrus [39]. No explanation, however, was given why particularly the rps16-trnQ and accD-psiA spacers had been chosen and not one of the highest ranked ones in terms of variability. The authors noted that the frequency of microstructural mutations in both spacers studied was markedly higher than of substitutions and that haplotypes were mostly defined by indels. Such a dominance of microstructural mutations over substitutions is typical of AT-rich sequence elements that constitute terminal stem-loops of introns and transcribed spacers which are often unique to small lineages of plants [11]. At the same time such sequence elements often exhibit high levels of homoplasly. Thus, the exclusive application of these elements to calculate networks or trees may potentially lead to wrong conclusions. Wuyun et al. [42] sequenced the rps16-trnQ and accD-psiA spacers to reconstruct a phylogenetic network of Pyrus ussuriensis in China, which was largely based on the presence or absence of indels in the two spacers. Compared with our results, the two regions used by Katayama et al. [47] and Wuyun et al. [48] are also not the most variable plastid regions in Pyrus: the trnQ-rps16 spacer ranks at place 24 for p-distances and at place 5 for PICs. The accD-psiA spacer ranks at place 18 for p-distances and at place 20 for PICs.

Plastid markers proposed for Pyrus

Four intergenic spacers of 900 to 1000 bp and the rpl16 group II intron (ca. 1000 bp) are proposed here to be sequenced for evolutionary studies in Pyrus (Table 4). They were selected from the most variable genomic regions (Table 3) considering an efficient sequencing strategy (see methods section).

Among the regions with a minimum size of 500 bp, the ndhC-trnV and trnR-accD spacers rank 3rd and 4th according to p-distances, and ndhC-trnV has the highest number of PICs. Both can be sequenced with just one primer (either forward or reverse). Thus, these spacers are especially useful if large sample numbers need to be analysed. The ndhC-rpl32 spacer (ranked 3rd of the regions >500 bp in Table 4) was not considered further because there are two large microsatellites. This fragment can therefore not be sequenced with two primers. The same problem occurs in the rps16-trnK spacer (ranked 4th of the regions >500 bp in Table 4) where two poly G and one poly T are likely to cause sequencing problems with pheograms unreadable after the homonucleotide stretches. The trnQ-rps16 and pshM-trnD spacers follow in the ranking. Both also have polyA/T microsatellites. While they can be covered with two primer reads that overlap at the microsatellite, they may not be as efficiently sequenced than the ndhC-trnV and trnR-accD spacers for large sample numbers. The rpl16 intron (ranked at 7th position of the regions >500 bp in Table 4) is particularly recommended because it was shown to also possess a high phylogenetic structure R in different angiosperm sequence data sets [43-45]. Multiple rpl16 sequence alignments can therefore be expected to yield well-resolved and well-supported trees also in Pyrus. The intron can be co-amplified with the rpl14-rps16 spacer. The use of the reverse primer PYR-rpl16R (Table 4) will allow to sequence the whole intron with one read. The rpl16 intron contains a polyA/T stretch of variable length in different species of Pyrus (see also Fig. 5c), what implies that an additional forward primer read may be necessary to cover the whole intron in some samples.

Primers were newly designed for trnR-accD as this region to our knowledge has never been used in any evolutionary study so far. For ndhC-trnV, primers were available [46] but we designed a new Pyrus-specific reverse primer in order to completely cover the spacer-exon boundary. For trnQ-rps16, the universal primers designed by Shaw & al. [9] work for Pyrus as well. Available primers for pshM-trnD [47] were re-designed for Pyrus to avoid mismatches in the forward and then to obtain a similar melting temperature in the reverse primer. For the rpl16 intron, primers were also adapted to Pyrus following the general amplification strategy of [43] and [44] with a forward primer that anneals to the rps3 exon. This ensures that the rpl16 intron can be amplified and sequenced completely. The universal reverse primer rpl16R [48] was replaced by a Pyrus-specific primer that anneals further downstream to cover the intron-exon boundary.

Comparison of plastid genomes with low p-distances in angiosperms. In addition to Pyrus, we explored variability patterns in plastid genome pairs of Oenothera argillacea and O. parviflora (Onagraceae), Olea europaea and O. woodiana (Olacaceae), and Cymbidium sinense and C. tortisepalum (Orchidaceae) which have comparable low p-distances (Table 2). The variability patterns of all four genome pairs are illustrated using a Circos-plot (Figs. 2-4). Each genome pair has different regions with highest p-distances and highest numbers of PICs, resulting in a genome pair-specific ranking (Table 3). The results of the pairwise comparisons of individual introns and spacers for each genome pair are provided in Table S2.

The SNPs and indels are almost evenly spread across the LSC and the SSCs in Olea. In Cymbidium, SNPs and indels are more clustered. The plastid genomes of Pyrus and Oenothera exhibit strong variation in certain areas, e.g. between trnT and rpoB (Figs. 1, 3) but alsoosallo homogeneously distributed mutations across their genomes. The Olea genome stands out by many more SNPs than indels, while the other genomes have almost as many indels as SNPs.
In our summary of the 30 most variable genomic regions including all four genome pairs, 77 different regions appear in total (Table 3). It is noteworthy that only two spacers, ndhF–rpl32 and trnK–rps16, are consistently placed among the 30 most variable regions. Eight spacers appear three times: atpD–rpl2, psaA–ycf3, psbR–psbT, rps4–trnT, trnQ–psbK, trnS–trnG, trnT–psbD, and trnT–trnL.

Earlier comparisons of plastid genomes in angiosperms for marker selection. In an approach to explore hitherto unused plastid regions as phylogenetic markers, Shaw et al. [9] in 2007 compared whole plastid genomes in a comprehensive way. They analysed genome pairs from three different lineages of angiosperms [Atropa and Nicotiana (Solanaceae) for the asterids, Lotus and Medicago (Fabaceae) for the rosids, and Oryza and Saccharum (Poaceae) for the monocots]. They found nine previously unexplored plastid regions with high levels of variation based on the numbers of PICs: rpl32–trnL, trnQ–rps16, ndhC–trnV, ndhF–rpl32, psbD–trnT, psbF–petA, rps16–trnK, atpD–atpH, and petL–psbE. As noted before, we were interested to compare the distance levels of these genomes to the genome pairs examined here, as we expected considerable differences. The p-distances were indeed much higher and are here calculated as examiners here, as we expected considerable differences. The

| Region          | Amplified fragment | Primer name     | Primer sequence                  | Reference                  |
|-----------------|--------------------|-----------------|----------------------------------|----------------------------|
| ndhC–trnV       | 900 bp             | ndhC–F          | TGCCAAAAAGGAAATAACAC             | Goodson et al. [46]         |
| trnR–atpA       | 1000 bp            | trnR–F          | GTCTAATGGAGACAGACAGG             | this study                  |
| psbM–trnD       | 1350 bp            | atpA–180R       | GAAACRAACGCTATCTGATTCC           | this study                  |
| trnQ–rps16      | 900 bp             | trnQ (UUG)      | GCTGGCAGATGGATAAGGC              | this study                  |
| rpl16 intron    | 1300 bp            | rps16x1         | GTTCATCCATCACATCGTT              | this study                  |

Table 4. Genomic regions proposed for evolutionary analyses in Pyrus and primers for their amplification.

doi:10.1371/journal.pone.0112998.t004

In summary, lineage specific differences in variability and phylogenetic utility of plastid genomic regions were reported in various cases in flowering plants although there was never any standardized comparative approach to better understand this issue. Moreover, none of the previous studies explicitly addressed phylogenetic signal as being different from similarity-based variability, or looked at any molecular evolutionary characteristics.

Molecular evolution and lineage specific variability of genomic regions. Lineage-specific differences in variability are often explained by patterns of molecular evolution. It has been exemplarily demonstrated for regions such as psbA–trnH [53] or trnL–trnF [54] that variability is strongly influenced by structural constraints. Empirical analysis of petD group II intron sequences has further shown that increased length correlates with increased AT strongly influenced by al constraints Empirical analysis of

In summary, some kind of sequence variability, while others additionally or solely reconstructed phylogenetic trees based on small taxon sets to assess the phylogenetic utility of these regions. A spectrum of 37 plastid loci was reported as “highly variable” in the studies cited above. Most commonly mentioned were rpl32–trnL (7x), trnQ–rps16 (5x), trnK–rps16 (4x), and ndhC–trnV (4x). Nevertheless, the question remains how representative the earlier pairwise genome comparisons are, and to what extent their conclusions are also valid for other families and genera of flowering plants.

Shaw et al. [8] assumed a high universality of their results. But Daniell et al. [52], who compared plastid genomes of Solanaceae, found spacers with higher sequence divergence not mentioned in [8]. Timme et al. [49] analysed Asteraceae and indicated that their ranking of most variable regions barely overlapped with the ranking of Shaw et al., and suspected that “each family or major lineage will most likely have a unique set of variable regions” [43]. Shaw et al. [9] in 2007 found no less than 11 new highly variable markers not considered in their 2005 study therefore pointed to the need of a test-wise screening of the “universal” regions to find the most suitable one for a given lineage. Likewise, Dong et al. [13] stated that markers useful for one group may not be useful for another and recommended evaluating markers in detail before selecting them for further use. With the aim of resolving the species tree in the huge genus Solanum, Sarkinen and George [14] found that the average amount of variable characters differs within subclades of the genus. In their view, the degree to which the utility of a marker can be extended to more inclusive clades would then also be clade-specific.

In summary, lineage specific differences in variability and phylogenetic utility of plastid genomic regions were reported in various cases in flowering plants although there was never any standardized comparative approach to better understand this issue. Moreover, none of the previous studies explicitly addressed phylogenetic signal as being different from similarity-based variability, or looked at any molecular evolutionary characteristics.
contents of three types of indels (left side) and around SNPs (right side) in intervals of increasing size of each of our genome pairs. AT content distributions are displayed in boxplots with the cross showing the mean and the thick line referring to the median. Respective boxplots arranged along the x-axis then depict maximum distances of the intervals in each direction of the SNP. Apart from rare exceptions the surroundings of SNPs are distinctly more AT-rich than the whole genome (Fig. 6), indicating that substitutions occur predominantly in AT-rich stretches. The AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.

The AT content is significantly increased in sequence elements affected by microstructural changes (Fig. 6), both in SSRs and in the non-SSR indels. The SSRs are generally AT-rich, so the AT contents of the consensus sequences are displayed as dotted lines. Looking at indels, considerable differences are apparent in the frequency of different kinds among the four plant lineages. In Olea, length-variable polyA/T stretches are most common. In Oenothera, all three kinds of indels occur with almost equal frequency, while in Cymbidium and Pyrus indels without a clear motif predominate.
conserved across all genera. They form the stem of the domain. Terminal parts of the domain such as the length-variable polyA-stretch in Oenothera have no structural constraints and therefore evolve rather freely. In Olea, there are two substitutions (indicated with ambiguity codes in the secondary structure) and one length variable polyA stretch. Again they occur in the terminal stem-loop and have no influence on the structure. The *rpl16* intron is more variable in Pyrus than in the other genome pairs. The polyT-stretch of Olea and Pyrus (beginning at position 10) is hypothesized as homologous in the alignment. But the predicted secondary structures (Fig 5c) show that this polyT stretch forms different secondary structures caused by the different adjacent sequence elements. In Olea, it forms a bulge but in Pyrus it forms a stem-element together with a complementary ‘AAAAACACAAAAAA’ motif [12,54].

**Sequence variability versus phylogenetic signal.** It is important to note that sequence variability as such does not necessarily correlate with the amount of hierarchical phylogenetic signal in a multiple sequence matrix. Thus, *p*-distances and PICs, which are both measures of sequence variability and describe the similarity of sequences, will not necessarily indicate the phylogenetically most informative regions. The phylogenetic utility of genomic regions depends on the distribution and kind of character state transformations throughout the evolutionary history of the sequences. Several statistics have been proposed to measure the hierarchical phylogenetic signal (referring to the phylogenetic structure in a data set) that take into account the

---

Figure 6. AT content of indels and areas around substitutions. Boxplot representation of the AT content in different types of indels (polyN, short sequence repeats (SSR) and other indels) on the left side and in areas with different sizes around all substitutions (SNPs) in the genome on the right side for a) Pyrus spinosa and *P. pyrifolia*, b) Cymbidium tortisepalum and *C. sinense*, c) Oenothera parviflora and *O. argillicola* and d) Olea europaea and *O. woodiana*. The cross in each boxplot indicates the mean of the distribution, the thick line refers to the median. The dotted line shows the AT content of the whole consensus sequence. doi:10.1371/journal.pone.0112998.g006
number of resolved nodes and the statistical support for these nodes \cite{58,59}. Specifically, the statistics $R$, $B$, and $C$, have been defined by Müller et al. \cite{59}. The most important one, $R$, measures the proportion of resolved clades and their support in a tree inferred from a given data set relative to the maximum possible resolution and support. If all nodes have maximum support, $R$ will get the value 1; if the phylogeny is completely unresolved (consists only of polytomies), $R$ will have the value 0.

The empirical evaluation of phylogenetic structure in a genomic region generally requires a multiple sequence alignment of a representatively sampled clade. From the datasets that have been evaluated in detail using the $R$ statistic \cite{44,45,59}, it is evident that at one hand higher variability often leads to more phylogenetic information (simply because there are more potentially informative characters). On the other hand, there are marked differences in the quality of hierarchical phylogenetic signal coming from the same number of variable positions in different kinds of genomic regions \cite{45}. These can be explained by different molecular evolutionary patterns. The general trend across angiosperms is that high phylogenetic structure is found in intergenic spacers and group I and II introns, but not in protein-coding genes except $mutK$. In our case of very closely related plastid genomes, the effects of multiple changes of the same site, eventually leading to saturation, or reversals, will probably not be very significant because these sequences are just starting to diverge. Nevertheless, it will be interesting to determine the phylogenetic structure in the top-ranked genomic regions in terms of variability once more extensive taxon sets will be available.

Moreover, highly variable regions will be needed to distinguish haplotypes (or species), even if they do not provide sufficient information about their phylogeny \cite{44}. If haplotypes are used in the sense of individual alleles, the pure variability is most important. However, AT-rich sequence elements (often in stem-loops) can be highly homoplastic with respect to the evolution of microstructural mutations \cite{60,61}. The most extreme causes of homoplasy are inversions \cite{62,63}. Therefore, especially those markers that contain a single AT-rich mutational hotspot should be tested for congruence in signal with other plastid markers. Haplotype analyses often only use one or two markers, but experiences from other studies that have successfully reconstructed evolutionary relationships among closely related species indicate that the combination of four or five regions will be needed. An increased number of characters increases resolution and support also in network analyses \cite{64,65}.

**Implications for plastid marker development in angiosperms.** About 20–30 plastid spacers and introns are regularly sequenced for phylogenetic and haplotype analyses, for which universal amplification primers exist. Also, considerable progress has been made during recent years in predicting phylogenetic utility from molecular evolutionary patterns, revealing differences in phylogenetic structure of genes, group I and group II introns, and intergenic spacers \cite{10–12,45,59}. In this way, markers with high versus low phylogenetic signal can be distinguished. For higher levels of genetic distance levels (e.g. distantly related species, genera, and families of flowering plants), a detailed evaluation of markers is therefore hardly necessary because sound predictions can be made. But is it worth to sequence whole plastid genomes when very closely related groups of species are to be studied?

Our comparison of genome pairs at comparable low distances shows that the mutational dynamics of plastid genomic regions may follow its own path in different lineages. While the variability in the respective unique sequence elements contributes the major proportion of the overall variability of a genomic region at that level, this contribution will be increasingly negligible at higher distance levels. The exploration of the plastid genome for the most variable and most suitable regions will therefore be a worthwhile investment when genetic distances are low.

It is of course possible to sequence all or at least most of the 30 promising plastid regions individually for a small taxon set in a given group. However, the effort needed is quite high. At least 60 individual fragments would need to be PCR-amplified and sequenced using many individual primers. Since only three to five loci are usually sequenced in evolutionary studies, a large part of these data would be wasted or deposited in GenBank as “unpublished”. The sequencing and assembly of whole plastid genomes is still laborious, especially if critical areas of low coverage or homonucleotide stretches are verified by Sanger sequencing. Often overlooked costs have to be considered as well: this includes higher requirements for IT hardware and much increased time for sequence assembly and data management compared to traditional sequencing. Still, sequencing a complete plastid genome has many benefits over many single-marker PCRs. First, the complete genome sequence ensures that all genomic regions can be considered for marker development. And second, generating complete genomes allows for using the genome sequence for other studies, so that data are added in a complementary way to build proper information sources for the respective lineages (e.g., for comparative genomics, primer design, detection of plastid microsatellites, or extraction of regions for phylogenetic studies).

We therefore conclude that whole plastid genome sequencing will remain a worthwhile approach for marker development in evolutionary studies of plants.

**Supporting Information**

Table S1 Primers for verification of sequence parts ambiguously read by the 454 sequencing.

(XLSX)

Table S2 Ranking of all regions for the four genome pairs.

(XLSX)

File S1 Pairwise alignment of the plastid genomes of *Pyrus spinosa* and *P. pyrifolia*.

(FASTA)

File S2 Pairwise alignment of the plastid genomes of *Cymbidium tortisepalum* and *C. sinense*.

(FASTA)

File S3 Pairwise alignment of the plastid genomes of *Oenothera parviflora* and *O. argillicola*.

(FASTA)

File S4 Pairwise alignment of the plastid genomes of *Olea woodiana* and *O. europaea*.

(FASTA)

**Acknowledgments**

The genomics work was done at the Berlin Center for Genomics in Biodiversity Research (BeGenDiv). We greatly acknowledge the lab assistance of Susan Mbedi and Virginia Duwe. We are also grateful to Felix Heeger for assistance with the genome assembly and Susan Wicke (Munster) for helpful advice on genome annotation.

This is publication number 11 of the Berlin Center for Genomics in Biodiversity Research (BeGenDiv). The study was carried out as part of the project “Developing tools for conserving the plant diversity of the Transcaucasus” funded by VolkswagenStiftung.
Author Contributions
Conceived and designed the experiments: NK TB. Performed the experiments: NK HTV LN. Analyzed the data: NK LN HTV MA TB.

References
1. McCauley DE. (1995) The use of chloroplast DNA polymorphism in studies of gene flow in plants. Trends Ecol Evol 10: 198–202.
2. Sang T, Crawford DJ, Stuessy TF. (1997) Chloroplast DNA phylogeny, reticulate evolution, and biogeography of Paeonia (Paeoniaceae). Am J Bot 84: 1120–1136.
3. Sang T, Donoghue MJ, Zhang DM. (1997) Evolution of alcohol dehydrogenase genes in peonies (Paeonia): Phylogenetic relationships of putative nonhybrid species. Mol Biol Evol 14: 994–1007.
4. Konuma A. (2009) Chloroplast (pro parte major). Euro+Med Plantbase - the information resource for Euro-Mediterranean plant diversity. Available: http://www.europlante.org/home.html. Accessed February 2014 February 19.
5. Cuihui G, Spoonberg SA. (2003) Pyrus. Flora of China. eFloras.org. 2006. Published on the Internet http://www.efloras.org. Missouri Botanical Garden, St. Louis, MO & Harvard University Herbaria, Cambridge, MA. pp. 173–179.
6. Akopian JA. (2007) Opividakh roda Pyrus L. (Rosaceae) v Armenii (On the Pyrus L. (Rosaceae) species in Armenia). J F Rast Staat Resuls Armeni 16: 15–26.
7. Fedorov AA. (1954) Rod grusha Pyrus L. (The genus pear Pyrus L.) In: Sokolov SJ, editor. Dessy i kenturalski SSSR [The trees and shrubs of the USSR]. Moskva, Leningrad: Izdatelstvo Akademii Nauk SSSR [National Science Academy of the U.S.S.R. publishing]. pp. 378–414.
8. Shaw J, Lackey EB, Beck JT, Farmer SN, Liu W, et al. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Am J Bot 92: 142–166.
9. Shaw J, Lackey EB, Schilling EE, Small RL. (2007) Comparison of whole chloroplast genome sequences to choose nondocking regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am J Bot 94: 273–289.
10. Kelchner SA. (2002) Group II introns as phylogenetic tools: Structure, function, and evolutionary constraints. Am J Bot 89: 1651–1669.
11. Borsch T, Quandt D. (2009) Mutational dynamics and phylogenetic utility of the petD gene in angiosperms: the tortoise and the hare IV. Am J Bot 96: 149–199.
12. Korotkova N, Schneider J, Quandt D, Worberg A, Zukta G, et al. (2009) Phylogeny of the eucladot order Malpighiales: analysis of a recalcitrant clade with sequences of the petD group II intron. Plant Syst Evol 282: 201–226.
13. Dong W, Liu J, Wang L, Zhou S. (2012) Highly variable chloroplast genome loci for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS ONE 7: e53071.
14. Särkkäinen T, George M. (2013) Predicting plastid marker variation: Can complete plastid genomes reliably identify species? PLoS ONE 8: e71696.
15. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in the flowering plant families Geraniaceae, Goodeniaceae, and Arctostaphylaceae. Mol Phylogenet Evol 45: 45–65.
16. Michel F, Umesono K, Ozeki H. (1989) Comparative and functional anatomy of the group II intron. Plant Syst Evol 282: 201–228.
17. Kronen H, Tachibana M, Iketani H, Zhang S-L, Uematsu C. (2012) Comparative analysis of chloroplast DNA in 101 species: physical map and gene localization. Theor Appl Genet 106: 303–310.
18. Katayama H, Tachibana M, Iketani H, Zhang S-L, Uematsu C. (2012) Comparative analysis of plastome sequence variation and development of plastidial CAPS markers in common potato and related Solanum species. Genet Resour Crop Evol 59: 419–430.
52. Daniell H, Lee SR, Grevich J, Sasaki C, Quesada-Vargas T, et al. (2006) Complete chloroplast genome sequences of Solanum bulbocastanum, Solanum lycopersicum and comparative analyses with other Solanaceae genomes. Theor Appl Genet 112: 1503–1518.

53. Štorchová H, Olson MS (2007) The architecture of the chloroplast psbA-trnH non-coding region in angiosperms. Plant Syst Evol 268: 235–256.

54. Quandt D, Muller K, Stech M, Frahm J-P, Frey W, et al. (2004) Molecular evolution of the chloroplast trnL-F region in land plants. In: Goffinet B, Hollowell V, Magill R, editors. Molecular Systematics of Bryophytes. St Louis: Missouri Botanical Garden Press. pp. 15–37.

55. Wicke S, Schäferhoff B, dePamphilis CW, Muller KF (2014) Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. Mol Biol Evol 31: 529–545.

56. Lehmann K, Schmidt U (2003) Group II introns: Structure and catalytic versatility of large natural ribozymes. Crit Rev Biochem Mol Biol 38: 249–303.

57. Pyle AM, Lambowitz AM (2006) Group II Introns: Ribozymes that splice RNA and invade DNA. In: Gesteland RF, Cech TR, Atkins JF, editors. The RNA World. 3rd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. pp. 449–505.

58. Kallersjo M, Farris JS, Chase MW, Bremer B, Fay MF, et al. (1998) Simultaneous parsimony jackknife analysis of 2538 rbcL DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant Syst Evol 215: 259–287.

59. Muller K, Borsch T, Hilu KW (2006) Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: Contrasting matK, trnT-F, and rbcL in basal angiosperms. Mol Phylogenet Evol 41: 99–117.

60. Tesfaye K, Borsch T, Govers K, Bekele E (2007) Characterization of Coffea chloroplast microsatellites and evidence for the recent divergence of C. arabica and C. engensis chloroplast genomes. Genome 50: 1112–1129.

61. Borsch T, Hilu KW, Wiersema JH, Lohne C, Bartlott W, et al. (2007) Phylogeny of Nymphon (Nymphaceae): Evidence from substitutions and microstructural changes in the chloroplast trnT-trnF region. Int J Plant Sci 168: 630–671.

62. Quandt D, Muller K, Huttunen S (2003) Characterisation of the chloroplast DNA psbl-H region and the influence of dyad symmetrical elements on phylogenetic reconstructions. Plant Biol 5: 400–410.

63. Whitlock BA, Hale AM, Groff PA (2010) Intraspecific inversions pose a challenge for the trnH-psbA plant DNA barcode. PLoS ONE 5: e11533.

64. Fior S, Li M, Ostblom B, Violta R, Hodges SA, et al. (2013) Spatiotemporal reconstruction of the Aquilegia rapid radiation through next-generation sequencing of rapidly evolving cpDNA regions. New Phytol 198: 579–592.

65. Erixon P, Ouchman B (2008) Reticulate or tree-like chloroplast DNA evolution in Sileneae (Caryophyllaceae)? Mol Phylogenet Evol 48: 313–323.

66. Greiner S, Wang X, Herrmann RG, Rauwolf U, Mayer K, et al. (2008) The complete nucleotide sequences of the 5 genetically distinct plastid genomes of Oenothera, subsection Oenothera: II. Sequence evaluation and plastome evolution. Nucleic Acids Res 36: 2366–2378.

67. Besnard G, Hernandez P, Khadari B, Dorado G, Savolainen V (2011) Genomic profiling of plastid DNA variation in the Mediterranean olive tree. BMC Plant Biol 11.