plants

MDPI

Article

Variation in Chloroplast Genome Size: Biological Phenomena and Technological Artifacts

Ante Turudić [1,2,*], Zlatko Liber [2,3], Martina Grdiša [1,2], Jernej Jakše [4], Filip Varga [1,2] and Zlatko Šatović [1,2]

[1] Centre of Excellence for Biodiversity and Molecular Plant Breeding (CoE CroP-BioDiv), Svetošimunska c. 25, 10000 Zagreb, Croatia
[2] Faculty of Agriculture, University of Zagreb, Svetošimunska c. 25, 10000 Zagreb, Croatia
[3] Faculty of Science, University of Zagreb, Marulićev trg 9a, 10000 Zagreb, Croatia
[4] Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, 1000 Ljubljana, Slovenia
* Correspondence: aturudic@agr.hr; Tel.: +385-91-3141592

Abstract: The development of bioinformatic solutions is guided by biological knowledge of the subject. In some cases, we use unambiguous biological models, while in others we rely on assumptions. A commonly used assumption for genomes is that related species have similar genome sequences. This is even more obvious in the case of chloroplast genomes due to their slow evolution. We investigated whether the lengths of complete chloroplast sequences are closely related to the taxonomic proximity of the species. The study was performed using all available *RefSeq* sequences from the asterid and rosid clades. In general, chloroplast length distributions are narrow at both the family and genus levels. In addition, clear biological explanations have already been reported for families and genera that exhibit particularly wide distributions. The main factors responsible for the length variations are parasitic life forms, IR loss, IR expansions and contractions, and polyphyly. However, the presence of outliers in the distribution at the genus level is a strong indication of possible inaccuracies in sequence assembly.

Keywords: genome databases; chloroplast genome; sequence length; taxonomy

Citation: Turudić, A.; Liber, Z.; Grdiša, M.; Jakše, J.; Varga, F.; Šatović, Z. Variation in Chloroplast Genome Size: Biological Phenomena and Technological Artifacts. *Plants* **2023**, *12*, 254. https://doi.org/10.3390/plants12020254

Academic Editors: Li'na Yin and Xiaomin Liu

Received: 4 November 2022
Revised: 31 December 2022
Accepted: 3 January 2023
Published: 5 January 2023

1. Introduction

Chloroplasts are cell organelles in which photosynthetic reactions occur. They have their own genome (cpDNA)—the plastome—which is generally described as circular [1], but in some cases has been described as a linear, multimeric circular or branched double-stranded molecule [2–4]. The plastome is usually 120 to 170 kbp long and consists of 120 to 130 genes (e.g., [5,6]).

The chloroplast genome sequence is a valuable source of data for evaluating plant evolution and taxonomy at different taxonomic levels [7]. Comparing the average chloroplast genome to the mitochondrial or nuclear genome, the gene composition is highly conserved with a collinear sequence arrangement. The slow evolution of most plastomes may be explained by the organization of chloroplast genes into operons, the mostly uniparental mode of inheritance, the activity of highly effective repair mechanisms, and very rare plastid recombination [8]. However, when there are major rearrangements in the chloroplast genome, they are usually associated with a parasitic and mycoheterotrophic lifestyle [9,10], and with an unusual mode of chloroplast inheritance, paternal or biparental [11]. Plant plastome size is specific to particular taxonomic groups such as order or family, but differences have very rarely been found between species of the same genus [12–15]. Three main factors are thought to be responsible for variation in chloroplast genome length: variation in intergenic regions (e.g., rice, family Pinaceae and genus *Oenothera*), variation in IR regions (e.g., gymnosperms, family Poaceae and Fabaceae), and gene loss (e.g., parasitic plants) [16].

However, there is no doubt that the cpDNA genomes of genetically close species are similar and that the conclusions derived should be valuable. There are numerous biological explanations for cases in which sequence similarity does not follow genetic proximity. Nevertheless, it should not be overlooked that some differences are due not to biology but to the technology used for sequence assembly. This is because the numerous different pipelines for genome sequence assembly do not always produce the same results. Therefore, observed differences between species of the same genus may indicate possible inaccuracies in the assembly process.

Advances in sequencing technologies have made genomic data more affordable. Together with the development of assembly methods, this has led to an increasing number of assembled genomes. This is clearly observed in the case of cpDNA data, as it is much easier to assemble chloroplast DNA than nuclear DNA. The reason for this is the higher copy number of cpDNA molecules [17,18] and the small size of the genome [1]. This is reflected in the number of genomes available in public databases. There is at least an order of magnitude more cpDNA than nuclear plant genomes in these databases [19]. At the time of writing, more than 10,000 plant cpDNA sequences were available in the NCBI (National Center for Biotechnology) Reference Sequence (*RefSeq*) collection, which provides a comprehensive, integrated, nonredundant, and well-annotated set of sequences. The number of cpDNA sequences in the NCBI *GenBank* database, which contains a collection of all publicly available DNA sequences, is a few times larger.

In general, bioinformatic solutions are developed based on biological knowledge and previous results on the subject; this is true for both chloroplast genome assemblers and annotation tools. Specialized chloroplast genome assemblers use existing results as reference sequences [20,21] or as starting positions for seed-and-extend assembly [22]. Therefore, assembly tools benefit from a larger pool of reference sequences from which to select the closest sequences, and the assembly process can be guided by their properties. Similarly, chloroplast annotation tools [23–27] rely on information from the gene database. In our opinion, the biological property of a slow evolutionary rate of the chloroplast genome combined with a large amount of available assembled sequence data may also be a fruitful combination to distinguish biological phenomena from technological artifacts.

We analyzed *RefSeq* chloroplast sequence length distribution using data from more than 5500 cpDNA sequences from plant species belonging to the asterid and rosid clades. The objectives of this study were to (a) assess the distribution of chloroplast sequence lengths at the family and genus levels, (b) identify families/genera with particularly wide distributions, and (c) detect outliers in the distribution. Possible explanations for wide distributions and the presence of outliers are provided. The utility of the results for the development of specialized chloroplast bioinformatics tools is discussed. We implemented a bioinformatics pipeline that can be used to analyze sequences of any taxa and store the result of each step in a concise format.

2. Results

2.1. Data Acquisition

A total of 5545 sequence summaries were acquired, of which 2534 and 3011 sequences were from the asterid and rosid clades, respectively (Table 1). The plant families with at least 20 sequences (the number chosen as the threshold) are shown in Figure 1. The data presented in the figure include 5076 sequences, representing 91.5% of the dataset, with asterid and rosid clades represented by 2337 and 2739 sequences, respectively. The largest families per clade were Asteraceae (asterids) and Fabaceae (rosids), with 543 and 480 sequences, respectively. In the NCBI Taxonomy Database, the total number of species per family ranged from 96 (Cornaceae) to 14,403 (Asteraceae) in the asterid clade and from 65 (Ulmaceae) to 12,978 (Fabaceae) in the rosid clade. Of all species in these families listed in the NCBI taxonomy, 5.35% had a *RefSeq* cpDNA sequence. For a comparison to the status of the same clade data from late 2021, see [28] (2022) for a similar presentation.

Table 1. Total number of taxa and sequences collected and analyzed. Summary of the results on the distribution range and outliers.

Parameter	Family Level			Genus Level		
	Asterids	**Rosids**	**Total**	**Asterids**	**Rosids**	**Total**
Total number of taxa	68	89	157	690	905	1595
No. of taxa analyzed	26	28	54	49	59	108
No. sequences in analyses	2337	2739	5076	1285	1410	2695
Minimum no. of sequences	21 Balsaminaceae	21 Ulmaceae	21	10 *Rhododendron*	10 *Glycine*	10
Maximum no. of sequences	543 Asteraceae	480 Fabaceae	543	178 *Solanum*	78 *Acer*	178
Examination of wide distributions						
No. and % of sequences IQR/median > 1%	11 (42.31%)	13 (46.43%)	24 (44.44%)	8 (16.33%)	6 (10.17%)	14 (12.96%)
No. and % of sequences IQR/median > 10%	3 (11.54%)	3 (10.71%)	6 (11.11%)	1 (2.04%)	3 (5.08%)	4 (3.70%)
Outlier detection						
No. of distributions with outliers	18 (69.23%)	22 (78.57%)	40 (74.07%)	31 (63.27%)	38 (64.41%)	69 (63.89%)
Total number of outliers	188 (8.04%)	129 (4.71%)	317 (6.25%)	103 (8.02%)	97 (6.88%)	200 (7.42%)

2.2. Distribution Assessment

We collected a summary of 5545 chloroplast sequences. Families with 20 or more sequences and genera with 10 or more sequences were used to calculate the distribution. The number of sequences used to calculate the length distribution and the statistics of the resulting boxplot values are shown in Table 1. The input and all result data are listed in Supplementary Table S1. The evaluation of the distributions was carried out for 54 families, which corresponds to 34.4% of all the families collected. These families contained 5076 sequences, which accounted for 91.5% of the total dataset. An assessment was performed for 108 genera, which contained 2695 sequences, representing 48.3% of the dataset. The number of sequences in the family records ranged from 21 to 543 sequences, and in the genera, it ranged from 10 to 178 sequences.

The histograms of the resulting IQR/median values for the records of genera and families are shown in Figure 2. In the case of families, 30 records (55.56%) have an IQR/median value of less than 1%; this proportion is less than 5.5% in all (88.89%) except six families (11.11%). The distributions of the genera are generally narrower. As seen in the histogram, 64 of the genus records (59.26%) have an IQR/median of less than 0.3%, and 94 of the records (87.04%) have a ratio value of less than 1%. The rest of the genus groups (12.96%) have ratios greater than 1%. The results show that the distributions become narrower and have fewer outliers when lower taxonomic rank (i.e., genus) is used, which is to be expected.

After examining the IQR/median ratio values obtained, we decided to use 1% for genera and 10% for families as a threshold to consider a distribution as wide. Wide distributions were detected in 6 cases (11.11%) for families and in 14 cases (12.96%) for genera. The total number of outliers detected in the family distributions was 317, representing 6.25% of the species included in the analysis, with 40 families (74.07%) containing outliers. At the genus level, 69 genera (63.89%) contained outliers, and there was a total of 200 outliers, corresponding to 7.42% of the sequences analyzed.

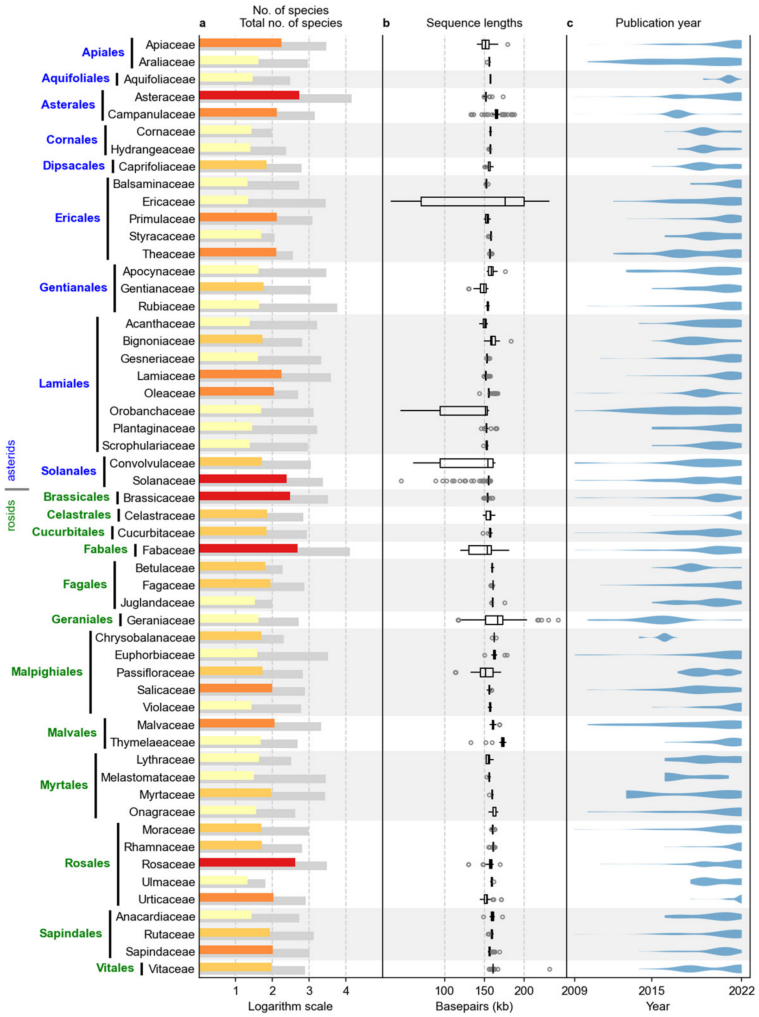

Figure 1. Number of *RefSeq* complete chloroplast sequences, distribution of sequence lengths, and publication years for the families of asterids and rosids containing 20 or more sequences. (**a**) Number of species according to NCBI taxonomy (gray bar) and number of available *RefSeq* chloroplast sequences (colored bar), where yellow represents 20–49, light orange 50–99, dark orange 100–199 and red ≥200 species. (**b**) Box plots of sequence lengths. (**c**) Violin plots of the number of published sequences by year.

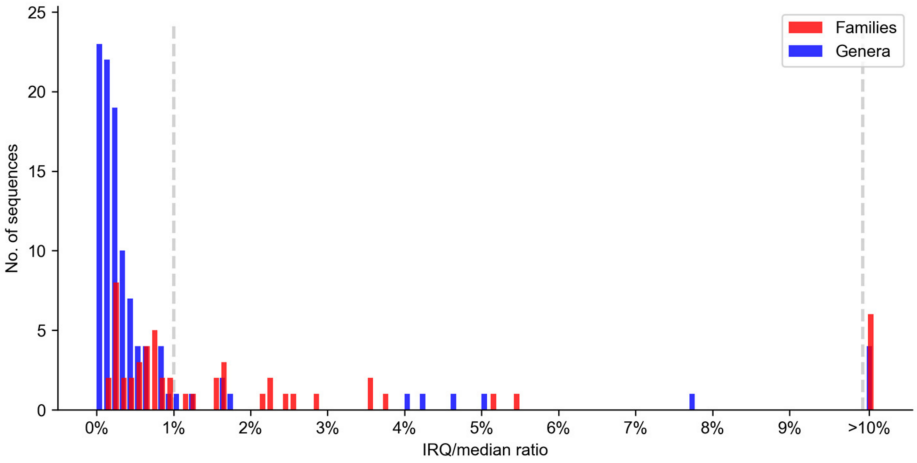

Figure 2. Histograms of IQR/median ratio values at the family (red) and genus levels (blue). Dashed vertical lines represent the thresholds used for family (10%) and genus (1%) levels.

According to the available sequences and the parameters studied (percentage of sequences with IQR/median > 1% and >10%, percentage of family/genus distributions with outliers, and percentage of sequences that were outliers at the family/genus level), both asterid and rosid clades provided fairly similar results (Table 1).

2.3. Examination of Wide Distributions

2.3.1. Family Level

Table 2 lists families with distributions considered wide, showing an IQR/median ratio greater than 10%. The table contains the basic data; further details are shown in Figure 3, where the distribution is shown with a box plot and decomposed into the largest six genera, whose distributions are shown with violin plots. The results used to create the table and figure can be found in Table S1, worksheet "*WideFamilies*".

Table 2. Families showing wide distributions of chloroplast genome lengths: number of genera with *RefSeq* sequences, number of sequences, IQR/median ratio, and list of genera with wide distributions.

Family	No. of Genera	No. of Sequences	IQR/Median Ratio	Genera Showing Wide Distributions
Ericaceae [a]	9	22	73.54%	*Rhododendron*
Convolvulaceae [a]	11	52	43.30%	*Cuscuta*
Orobanchaceae [a]	22	49	39.07%	-
Fabaceae [r]	223	480	18.53%	*Medicago*, *Lathyrus*
Geraniaceae [r]	5	42	12.95%	*Pelargonium*, *Erodium*
Passifloraceae [r]	4	53	10.60%	*Passiflora*

[a] asterid clade. [r] rosid clade.

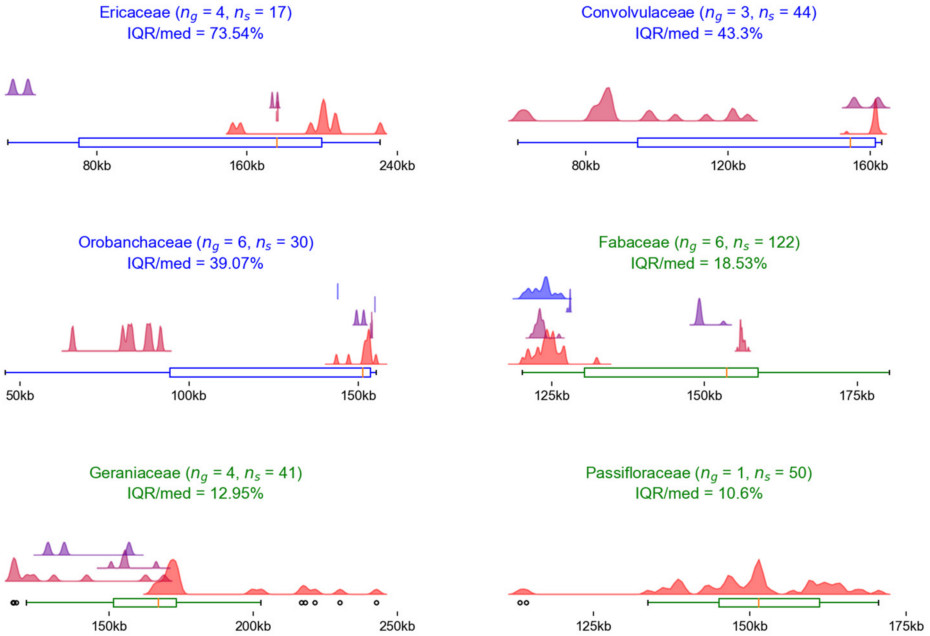

Figure 3. Box plots of the length distributions of the chloroplast genomes in the six families with the widest distributions. The length distributions of the largest genera within the families are represented by violin plots in different colors. A maximum of six of the largest genera with a number of sequences greater than 1 are shown. The numbers in brackets show the number of genera in the family (n_g) and the number of sequences in the genera presented (n_s). The title shows the IQR/median ratio for the family. The title and color of the boxplot represent the family clade: asterids (blue) or rosids (green). The numbers are arranged in descending order of IQR/median ratio values.

Three of the six families with the widest chloroplast genome length distributions were those for which loss of inverted repeats (IR) was reported in some taxa: Ericaceae [29,30], Fabaceae [31,32], and Geraniaceae [15,33].

The occurrence of parasitic taxa was reported in two families with wide distributions. The Convolvulaceae family contains a single parasitic genus, *Cuscuta* [34,35], which has much shorter sequences (60–125 kbp) than those of the other Convolvulaceae genera (153–162 kbp). The Orobanchaceae family is considered the largest predominantly parasitic angiosperm family [36]. The results show that the sequence lengths of the Orobanchaceae family vary among genera (e.g., *Phelipanche* 62–63 kbp, *Orobanche* 65–91 kbp, *Cistanche* 94–102 kbp, *Aphyllon* 107–121 kbp, *Pedicularis* 143–155 kbp, *Rehmannia* 153–154 kbp), which is because this family contains both parasitic and hemiparasitic species.

The dataset of the family Passifloraceae contained mainly species of the genus *Passiflora*, in which extensive genomic alterations were detected, including inversions, gene and intron losses, and several independent IR expansions and contractions [37,38].

2.3.2. Genus Level

The distributions of genera with IQR/median ratios greater than 1% are shown in Figure 3. The distribution is shown with a combined box and violin plot. The results used to create the figure can be found in Table S1, "*WideGenera*" worksheet.

For all genera showing a wide distribution of sequence lengths, some specific reasons for the variation in chloroplast genome length have been previously reported (Table 3). These reasons were the following: parasitic life form (e.g., *Cuscuta*), IR loss (e.g., *Erodium*, *Lathyrus*, *Medicago*, *Rhododendron*), IR expansions and contractions (e.g., *Passiflora*, *Pelargonium*), and polyphyly (e.g., *Amphilophium*, *Gentiana*, *Euphorbia*, *Lobelia*, *Peucedanum*, *Primula*, *Seseli*). The parasitic genus *Cuscuta* had the widest distribution of chloroplast genome length, ranging from 60 to 125 kbp. The results showed that three *Cuscuta* species (*C. exaltata*, *C. japonica*, and *C. reflexa*) with the longest sequences were hemiparasitic and belonged to the same paraphyletic group [35]. For the remaining genera, it is noteworthy that the IR changes seem to affect the distribution of sequence lengths more than the reported polyphyly of the genera in question. In addition, it is expected that the resolution of generic polyphyly will lead to the creation of new genera whose distribution will be much narrower.

Table 3. Genera showing wide distributions of chloroplast genome lengths: family, number of sequences, IQR/median ratio, and reported factor responsible for variation in chloroplast genome length.

Genus	Family	No. of Sequences	IQR/Median Ratio	Reported Factor
Cuscuta	Convolvulaceae [a]	20	19.44%	Parasitic life form [34,35]
Erodium	Geraniaceae [r]	10	18.12%	IR loss [15,33]
Pelargonium	Geraniaceae [r]	22	18.06%	IR expansions [15,39–41]
Passiflora	Passifloraceae [r]	50	10.34%	IR expansions and contractions [37,38]
Gentiana	Gentianaceae [a]	32	7.71%	Polyphyly [42]
Rhododendron	Ericaceae [a]	10	5.04%	IR loss [29,30]
Seseli	Apiaceae [a]	12	4.96%	Polyphyly [43]
Peucedanum	Apiaceae [a]	10	4.28%	Polyphyly [43]
Amphilophium	Bignoniaceae [a]	11	4.09%	Polyphyly [44]
Medicago	Fabaceae [r]	31	1.76%	IR loss [31,32]
Lathyrus	Fabaceae [r]	13	1.68%	IR loss [31,32]
Primula	Primulaceae [a]	82	1.61%	Polyphyly [45–47]
Euphorbia	Euphorbiaceae [r]	20	1.26%	Polyphyly [48,49]
Lobelia	Campanulaceae [a]	53	1.08%	Polyphyly [50]

[a] asterid clade. [r] rosid clade.

Apart from the fact that some genera showed a wide distribution, which can be explained by the biological phenomena mentioned above, some possible outliers were also detected, which can be clearly observed in Figure 4.

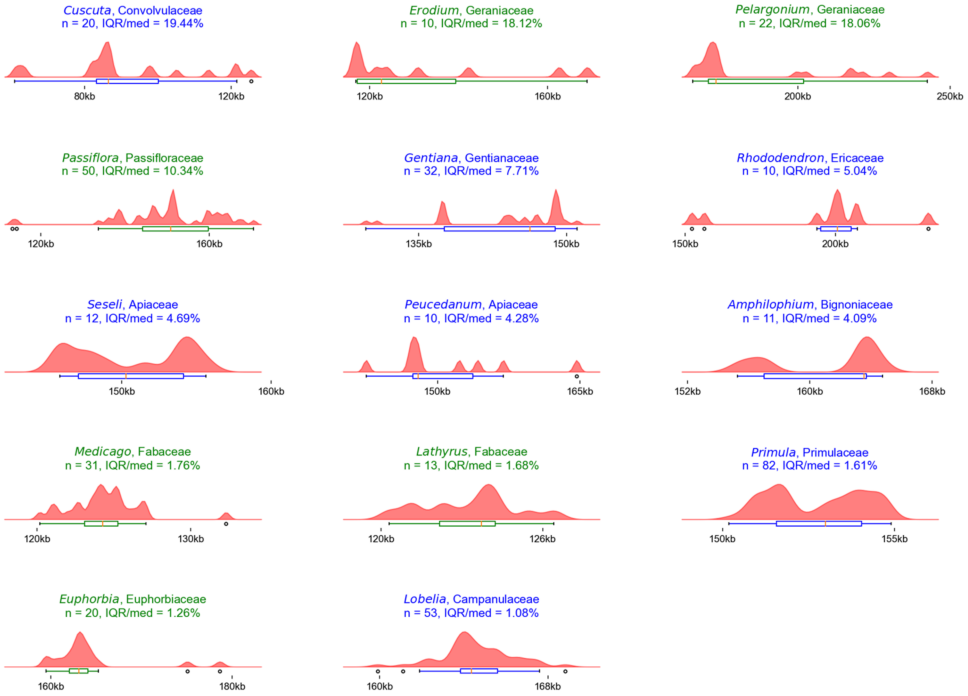

Figure 4. Composite box and violin plots of the length distributions of chloroplast genomes within genera showing wide distributions. The titles of the plots show the name of the genus and family, as well as the values for the number of sequences in the genus (n) and the IQR/median ratio of the distribution. The title and color of the boxplot represent the family clade: asterids (blue) or rosids (green). The genera are arranged in descending order of IQR/median ratio.

2.4. Outlier Detection at the Genus Level

Outliers were deetcted in 69 genera, representing 63.89% of the records where the length distributions were evaluated. Two hundred outliers were identified, which corresponded to 7.42% of the sequences used for the evaluation. For these species, we downloaded a summary of all available chloroplast sequences from the *GenBank* database. A total of 613 *GenBank* sequence summaries were acquired.

Note that the *GenBank* database contains both *RefSeq* sequences and submissions of the same sequences before they were included in the *RefSeq* database. Thus, 400 of the 613 sequences that we originally collected in the *RefSeq* database were identical sequences. These 213 alternative sequences belonged to 80 different species. Among them, we detected chloroplast sequences of 57 species (71.25%) whose length was closer to the median of the genus distribution than the original *RefSeq* sequence. A selection of cases where alternative *GenBank* sequences fit the genus distribution better than the original *RefSeq* sequence is provided in Table 4.

Table 4. Examples of *RefSeq* sequences where alternative *GenBank* sequences were detected that are closer to the median of the genus distribution. The accession number, sequence length, and publication date of the sequences are provided.

Species	Genus Median	*RefSeq* Sequence			Alternative Sequence		
		Accession	Length	Publication Date	Accession	Length	Publication Date
Angelica sinensis	146,962	NC_042826	142,485	25 June 2019	MW820164	146,952	5 September 2021
Ficus auriculata	160,363	NC_053837	162,558	26 March 2021	MZ662866	160,361	31 August 2022
Fragaria mandshurica	155,621	NC_018767	129,805	14 October 2012	MW537846	155,640	30 March 2022
Fragaria vesca	155,621	NC_018766	129,788	14 October 2012	KC507757	155,620	26 July 2016
Vitis romanetii	160,971	NC_056348	232,020	20 June 2021	MW592524	160,976	16 March 2022

In genera with wide distributions, outliers were detected in seven datasets (Figure 4), with a total of 13 outlier sequences. Alternative sequences were detected for two species, both closer to the median of the genus distribution than the original sequences. Details of the results can be found in Table S1, worksheets "*Outliers*" and "*Alternatives*".

3. Discussion

Our survey was prompted by the consideration that closely related species are more likely to share similar genome sizes and characteristics [16,51] and that a large amount of available cpDNA sequence data can be used to improve bioinformatic solutions for chloroplast assembly and annotation. In developing a method to quantify the expected proximity of related cpDNA sequences, the method must take into account potential differences arising from the technology used for sequence assembly in addition to the biological differences we wish to measure.

Given the large amount of data available, dissimilarities due to technology could be detected. These dissimilarities can range from problems of uniformity [52,53] to possible errors of assembly. Standardization problems, when discovered, can be resolved relatively easily by reformatting and reannotating the sequence data. On the other hand, sequence assembly inaccuracies are difficult to detect and virtually impossible to demonstrate without creating a new sequence assembly based on newer and presumably more accurate technologies and tools.

The problem of using existing data and the difficulty of detecting inadequacies in sequencing are related. To solve them, we need a quantification or model for the relationship between taxonomic proximity and differences in cpDNA sequences or some genomic characteristics. By default, genomic differences were detected by the alignment of sequences. In the case of cpDNA, this is not straightforward because the genome is circular and usually contains structures induced by inverted repeats (IRs). Therefore, the sequence format should be standardized [53]. However, a reliable IR identification method is still not available [28], while a different standardization approach would be required for the IR loss clades, probably based on conserved gene loci.

The study of the relationship between taxonomic proximity and differences in cpDNA sequence lengths was performed at two taxonomic levels for families and genera. Generally, results at the lower taxonomic level of genus are used by bioinformatics tools. The families studied mostly had narrow length distributions. The most extreme distributions were detected in parasitic families or those with IR losses or changes. Length distributions in the genera were generally very narrow, with more than half of the cases having an IQR/median ratio of less than 0.25%. This means that more than half of the genus sequences had lengths less than 200 base pairs from the median of the genus distribution. Only 12.96% of the genera had wide distributions as defined (IQR/median ratio > 1%). For all these wide distributions, there are some clear biological explanations, which are the main factors for the length variations: parasitic life form IR loss, IR expansions and contractions, and polyphyly.

Distribution outliers were also detected and tested with data from the *GenBank* database. In most cases (71.25%), the alternative sequences of the same species matched the expected distributions of the genera better than the original outlier sequences. This suggests that outlier detection is a promising method to identify suspect assemblies that require closer examination and possible correction.

We presented a strategy for using large amounts of existing cpDNA sequence data to derive a useful quantification of the relationship between species closeness and sequence length. The sequence length is well conserved at the genus level. Specifically, for chloroplast genomes, the property is also conserved at higher taxonomic levels, except in cases of clades with known high genome loss (IRs loss or alterations, parasitic clades). These results are useful for the development of specialized chloroplast bioinformatics tools. In addition, we have demonstrated a simple and promising method for identifying imperfect assemblies by combining outlier detection with checking data against sequences from other databases.

4. Materials and Methods

4.1. Data Acquisition

Summaries of chloroplast genome sequences were downloaded from the NCBI *RefSeq* database (https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/ (accessed on 22 October 2022)). We acquired summaries of all sequences available at that time for asterid and rosid clades. These clades were chosen as examples because they contain a sufficient number of sequences to demonstrate the utility of the bioinformatics pipeline we present. The sequence summaries included information on species name, accession number, genome length, and publication date. A total of 5545 sequence summaries were acquired. Additionally, for outlier species, summaries of alternative complete chloroplast sequences were downloaded from the NCBI *GenBank* database (https://www.ncbi.nlm.nih.gov/nucleotide (accessed on 22 October 2022)). Taxonomic data were retrieved from the NCBI Taxonomy Database (https://www.ncbi.nlm.nih.gov/taxonomy (accessed on 22 October 2022)).

4.2. Bioinformatics Pipeline

The research was based on the characteristics of the complete chloroplast sequence length distribution within a family or a genus. We used the descriptive statistics of box plots to represent the distributions. A box plot is a method of describing a dataset using five values: the median of the sample, the first and third quartiles, and the lower and upper whiskers. The whiskers are 1.5× interquartile ranges (IQRs), and the IQR is the distance between the upper and lower quartiles. Values that fell outside the whiskers were considered outliers [54]. As the main descriptor for the width of the distribution, we used the ratio of IQR and median values (IQR/median), which were presented as percentages.

We implemented a bioinformatics pipeline to acquire the data we need and store all the research results. The tool is controlled by arguments that determine the scope of data to be analyzed and the thresholds used in the calculations. The results of an analysis are saved in an Excel spreadsheet, and graphs based on the results obtained can be generated.

The analysis is carried out in four steps: data acquisition, distribution assessment, examination of the wide distributions, and outlier detection. Data acquisition was performed by querying the *RefSeq* database for summaries of the complete chloroplast sequences of selected species. The downloaded summary data are stored in the Excel worksheet "*RefSeq*". For this research, we used sequence data from the asterid and rosid clades.

The assessment of the length distributions of chloroplast genomes is performed within families and genera containing at least a certain number of sequences. The calculated distribution characteristics are stored in the Excel worksheet "*Distributions*". In this research, we carried out an assessment for families with at least 20 sequences and genera with at least 10 sequences.

When examining taxa that have a wide distribution of chloroplast genome lengths, families and genera whose IQR/median value is above the specified threshold are filtered out. The filtered data are stored in the Excel worksheets "*WideAll*", "*WideFamilies*", and "*WideGenera*". For families, the assessment data also include the distribution characteristics of their genera to check their influence on the family distribution. In our research, we chose a threshold of 1% for the distributions of genera and 10% for the distributions of families based on the obtained IQR/median values.

The outlier test detects sequences that are outliers in the distribution of their genera. For species that are considered outliers, the tool downloads a summary of all complete chloroplast sequences from the *GenBank* database. It then tests whether there is a sequence between those whose length is closer to the median of the genus distribution than that of the *RefSeq* sequence. The list of all detected outliers from the genus distribution is stored in the Excel worksheet "*Outliers*", and the list of outliers where an alternative sequence is detected that is closer to the median of the distribution is saved in the worksheet "*Alternatives*". Note that the *RefSeq* database is part of the *GenBank* database, and the *GenBank* database

also contains versions of *RefSeq* sequences before they were included in *RefSeq*, so the query retrieves a summary for at least two sequences.

The pipeline to perform the analysis was implemented in Python using the Biopython package [55] to retrieve sequence summaries. Taxonomy analysis was performed using the ETE 3 [56] Python library. Figures were generated using the Python library Matplotlib [57]. The code is maintained in a public repository (https://github.com/CroP-BioDiv/cpdna_survey; accessed on 22 October 2022). The script receives arguments for the taxa to be analyzed, the minimum number of sequences to calculate the distribution, and the thresholds for the IQR/median ratio.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/plants12020254/s1, Table S1: Excel file with worksheets for input data and all analyses results.

Author Contributions: Conceptualization, A.T., Z.L., J.J. and Z.Š.; methodology, A.T., J.J. and Z.Š.; software, A.T.; validation, Z.L., J.J. and Z.Š.; formal analysis, A.T. and M.G.; resources, F.V., Z.L. and M.G.; data curation, A.T., Z.L., M.G. and F.V.; writing—original draft preparation, A.T.; writing—review and editing, Z.L., J.J., F.V. and Z.Š.; visualization, A.T. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project KK.01.1.1.01.0005 Biodiversity and Molecular Plant Breeding, at the Centre of Excellence for Biodiversity and Molecular Plant Breeding (CoE CroP-BioDiv), Zagreb, Croatia.

Data Availability Statement: The data that support the findings of this study are openly available in the NCBI Nucleotide Database (https://www.ncbi.nlm.nih.gov/nucleotide (accessed on 22 October 2022)) and NCBI Taxonomy Database (https://www.ncbi.nlm.nih.gov/taxonomy (accessed on 22 October 2022)).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jansen, R.K.; Ruhlman, T.A. Plastid Genomes of Seed Plants. *Photosynthesis* **2012**, *35*, 103–126.
2. Deng, X.-W.; Wing, R.A.; Gruissem, W. The Chloroplast Genome Exists in Multimeric Forms. *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 4156–4160. [CrossRef] [PubMed]
3. Bendich, A.J.; Smith, S.B. Moving Pictures and Pulsed-Field Gel Electrophoresis Show Linear DNA Molecules from Chloroplasts and Mitochondria. *Curr. Genet.* **1990**, *17*, 421–425. [CrossRef]
4. Bendich, A.J. Circular Chloroplast Chromosomes: The Grand Illusion. *Plant Cell* **2004**, *16*, 1661–1666. [CrossRef] [PubMed]
5. Ohyama, K.; Fukuzawa, H.; Kohchi, T.; Shirai, H.; Sano, T.; Sano, S.; Umesono, K.; Shiki, Y.; Takeuchi, M.; Chang, Z.; et al. Chloroplast Gene Organization Deduced from Complete Sequence of Liverwort Marchantia Polymorpha Chloroplast DNA. *Nature* **1986**, *322*, 572–574. [CrossRef]
6. Shinozaki, K.; Ohme, M.; Tanaka, M.; Wakasugi, T.; Hayashida, N.; Matsubayashi, T.; Zaita, N.; Chunwongse, J.; Obokata, J.; Yamaguchi-Shinozaki, K.; et al. The Complete Nucleotide Sequence of the Tobacco Chloroplast Genome: Its Gene Organization and Expression. *EMBO J.* **1986**, *5*, 2043–2049. [CrossRef]
7. Gielly, L.; Taberlet, P. The Use of Chloroplast DNA to Resolve Plant Phylogenies: Noncoding versus RbcL Sequences. *Mol. Biol. Evol.* **1994**, *11*, 769–777. [CrossRef]
8. Ruhlman, T.A.; Jansen, R.K. The Plastid Genomes of Flowering Plants. *Methods Mol. Biol.* **2014**, *1132*, 3–38. [CrossRef]
9. Bellot, S.; Renner, S.S. The Plastomes of Two Species in the Endoparasite Genus Pilostyles (*Apodanthaceae*) Each Retain Just Five or Six Possibly Functional Genes. *Genome Biol. Evol.* **2015**, *8*, 189–201. [CrossRef]
10. Roquet, C.; Coissac, É.; Cruaud, C.; Boleda, M.; Boyer, F.; Alberti, A.; Gielly, L.; Taberlet, P.; Thuiller, W.; Van Es, J.; et al. Understanding the Evolution of Holoparasitic Plants: The Complete Plastid Genome of the Holoparasite *Cytinus hypocistis* (*Cytinaceae*). *Ann. Bot.* **2016**, *118*, 885–896. [CrossRef]
11. Wicke, S.; Schneeweiss, G.M.; de Pamphilis, C.W.; Müller, K.F.; Quandt, D. The Evolution of the Plastid Chromosome in Land Plants: Gene Content, Gene Order, Gene Function. *Plant Mol. Biol.* **2011**, *76*, 273–297. [CrossRef]
12. Downie, S.R.; Palmer, J.D. Restriction Site Mapping of the Chloroplast DNA Inverted Repeat: A Molecular Phylogeny of the Asteridae. *Ann. Mo. Bot. Gard.* **1992**, *79*, 266. [CrossRef]
13. Goulding, S.E.; Olmstead, R.G.; Morden, C.W.; Wolfe, K.H. Ebb and Flow of the Chloroplast Inverted Repeat. *Mol. Gen. Genet.* **1996**, *252*, 195–206. [CrossRef]
14. Plunkett, G.M.; Downie, S.R. Expansion and Contraction of the Chloroplast Inverted Repeat in *Apiaceae subfamily Apioideae*. *Syst. Bot.* **2000**, *25*, 648. [CrossRef]

15. Guisinger, M.M.; Kuehl, J.V.; Boore, J.L.; Jansen, R.K. Extreme Reconfiguration of Plastid Genomes in the Angiosperm Family Geraniaceae: Rearrangements, Repeats, and Codon Usage. *Mol. Biol. Evol.* **2011**, *28*, 583–600. [CrossRef]
16. Xiao-Ming, Z.; Junrui, W.; Li, F.; Sha, L.; Hongbo, P.; Lan, Q.; Jing, L.; Yan, S.; Weihua, Q.; Lifang, Z.; et al. Inferring the Evolutionary Mechanism of the Chloroplast Genome Size by Comparing Whole-Chloroplast Genome Sequences in Seed Plants. *Sci. Rep.* **2017**, *7*, 1555. [CrossRef]
17. Bendich, A.J. Why Do Chloroplasts and Mitochondria Contain so Many Copies of Their Genome? *BioEssays* **1987**, *6*, 279–282. [CrossRef]
18. Heinhorst, S.; Cannon, G.C. DNA Replication in Chloroplasts. *J. Cell Sci.* **1993**, *104*, 1–9. [CrossRef]
19. Marks, R.A.; Hotaling, S.; Frandsen, P.B.; VanBuren, R. Representation and Participation across 20 Years of Plant Genome Sequencing. *Nat. Plants* **2021**, *7*, 1571–1578. [CrossRef]
20. Jin, J.J.; Yu, W.B.; Yang, J.B.; Song, Y.; Depamphilis, C.W.; Yi, T.S.; Li, D.Z. Get Organelle: A Fast and Versatile Toolkit for Accurate de Novo Assembly of Organelle Genomes. *Genome Biol.* **2020**, *21*, 241. [CrossRef]
21. McKain, M.R.; Wilson, M. Fast-Plast: Rapid de Novo Assembly and Finishing for Whole Chloroplast Genomes. Available online: https://github.com/mrmckain/Fast-Plast (accessed on 22 October 2022).
22. Dierckxsens, N.; Mardulyn, P.; Smits, G. NOVOPlasty: De Novo Assembly of Organelle Genomes from Whole Genome Data. *Nucleic Acids Res.* **2017**, *45*, e18. [CrossRef]
23. Zhong, X. Assembly, Annotation and Analysis of Chloroplast Genomes. 2020. Available online: https://research-repository.uwa.edu.au/en/publications/assembly-annotation-and-analysis-of-chloroplast-genomes (accessed on 22 October 2022).
24. Zheng, S.; Poczai, P.; Hyvönen, J.; Tang, J.; Amiryousefi, A. Chloroplot: An Online Program for the Versatile Plotting of Organelle Genomes. *Front. Genet.* **2020**, *11*, 576124. [CrossRef]
25. Tillich, M.; Lehwark, P.; Pellizzer, T.; Ulbricht-Jones, E.S.; Fischer, A.; Bock, R.; Greiner, S. GeSeq—Versatile and Accurate Annotation of Organelle Genomes. *Nucleic Acids Res.* **2017**, *45*, W6–W11. [CrossRef] [PubMed]
26. Qu, X.J.; Moore, M.J.; Li, D.Z.; Yi, T.S. PGA: A Software Package for Rapid, Accurate, and Flexible Batch Annotation of Plastomes. *Plant Methods* **2019**, *15*, 50. [CrossRef] [PubMed]
27. Huang, D.I.; Cronk, Q.C.B. Plann: A Command-Line Application for Annotating Plastome Sequences. *Appl. Plant Sci.* **2015**, *3*, 1500026. [CrossRef]
28. Turudić, A.; Liber, Z.; Grdiša, M.; Jakše, J.; Varga, F.; Šatović, Z. Chloroplast Genome Annotation Tools: Prolegomena to the Identification of Inverted Repeats. *Int. J. Mol. Sci* **2022**, *2022*, 10804. [CrossRef] [PubMed]
29. Fajardo, D.; Senalik, D.; Ames, M.; Zhu, H.; Steffan, S.A.; Harbut, R.; Polashock, J.; Vorsa, N.; Gillespie, E.; Kron, K.; et al. Complete Plastid Genome Sequence of Vaccinium Macrocarpon: Structure, Gene Content, and Rearrangements Revealed by next Generation Sequencing. *Tree Genet. Genomes* **2013**, *9*, 489–498. [CrossRef]
30. Martínez-Alberola, F.; Del Campo, E.M.; Lázaro-Gimeno, D.; Mezquita-Claramonte, S.; Molins, A.; Mateu-Andrés, I.; Pedrola-Monfort, J.; Casano, L.M.; Barreno, E. Balanced Gene Losses, Duplications and Intensive Rearrangements Led to an Unusual Regularly Sized Genome in Arbutus Unedo Chloroplasts. *PLoS ONE* **2013**, *8*, e79685. [CrossRef]
31. Wojciechowski, M.F.; Sanderson, M.J.; Hu, J.M. Evidence on the Monophyly of *Astragalus* (*Fabaceae*) and Its Major Subgroups Based on Nuclear Ribosomal DNA ITS and Chloroplast DNA TrnL Intron Data. *Syst. Bot.* **1999**, *24*, 409. [CrossRef]
32. Wojciechowski, M.F.; Sanderson, M.J.; Steele, K.P.; Liston, A. Molecular Phylogeny of the "Temperate Herbaceous Tribes" of Papilionoid Legumes: A Supertree Approach. *Adv. Legume Syst.* **2000**, *9*, 277–298.
33. Blazier, C.C.; Guisinger, M.M.; Jansen, R.K. Recent Loss of Plastid-Encoded Ndh Genes within Erodium (Geraniaceae). *Plant Mol. Biol.* **2011**, *76*, 263–272. [CrossRef]
34. Stefanović, S.; Olmstead, R.G. Testing the Phylogenetic Position of a Parasitic Plant (*Cuscuta*, *Convolvulaceae*, *Asteridae*): Bayesian Inference and the Parametric Bootstrap on Data Drawn from Three Genomes. *Syst. Biol.* **2004**, *53*, 384–399. [CrossRef]
35. Park, I.; Song, J.H.; Yang, S.; Kim, W.J.; Choi, G.; Moon, B.C. Cuscuta Species Identification Based on the Morphology of Reproductive Organs and Complete Chloroplast Genome Sequences. *Int. J. Mol. Sci.* **2019**, *20*, 2726. [CrossRef]
36. Li, X.; Zhang, T.C.; Qiao, Q.; Ren, Z.; Zhao, J.; Yonezawa, T.; Hasegawa, M.; Crabbe, M.J.C.; Li, J.; Zhong, Y. Complete Chloroplast Genome Sequence of Holoparasite *Cistanche deserticola* (*Orobanchaceae*) Reveals Gene Loss and Horizontal Gene Transfer from Its Host *Haloxylon ammodendron* (*Chenopodiaceae*). *PLoS ONE* **2013**, *8*, e58747. [CrossRef]
37. Rabah, S.O.; Shrestha, B.; Hajrah, N.H.; Sabir, M.J.; Alharby, H.F.; Sabir, M.J.; Alhebshi, A.M.; Sabir, J.S.M.; Gilbert, L.E.; Ruhlman, T.A.; et al. Passiflora Plastome Sequencing Reveals Widespread Genomic Rearrangements. *J. Syst. Evol.* **2019**, *57*, 1–14. [CrossRef]
38. Cauz-Santos, L.A.; da Costa, Z.P.; Callot, C.; Cauet, S.; Zucchi, M.I.; Bergès, H.; van den Berg, C.; Vieira, M.L.C. A Repertory of Rearrangements and the Loss of an Inverted Repeat Region in Passiflora Chloroplast Genomes. *Genome Biol. Evol.* **2020**, *12*, 1841–1857. [CrossRef]
39. Chumley, T.W.; Palmer, J.D.; Mower, J.P.; Fourcade, H.M.; Calie, P.J.; Boore, J.L.; Jansen, R.K. The Complete Chloroplast Genome Sequence of Pelargonium × Hortorum: Organization and Evolution of the Largest and Most Highly Rearranged Chloroplast Genome of Land Plants. *Mol. Biol. Evol.* **2006**, *23*, 2175–2190. [CrossRef]
40. Röschenbleck, J.; Wicke, S.; Weinl, S.; Kudla, J.; Müller, K.F. Genus-Wide Screening Reveals Four Distinct Types of Structural Plastid Genome Organization in *Pelargonium* (*Geraniaceae*). *Genome Biol. Evol.* **2017**, *9*, 64–76. [CrossRef]
41. Weng, M.L.; Ruhlman, T.A.; Jansen, R.K. Expansion of Inverted Repeat Does not Decrease Substitution Rates in Pelargonium Plastid Genomes. *New Phytol.* **2017**, *214*, 842–851. [CrossRef]

42. Favre, A.; Pringle, J.S.; Heckenhauer, J.; Kozuharova, E.; Gao, Q.; Lemmon, E.M.; Lemmon, A.R.; Sun, H.; Tkach, N.; Gebauer, S.; et al. Phylogenetic Relationships and Sectional Delineation within *Gentiana* (*Gentianaceae*). *Taxon* **2020**, *69*, 1221–1238. [CrossRef]
43. Spalik, K.; Reduron, J.P.; Downie, S.R. The Phylogenetic Position of *Peucedanum sensu* Lato and Allied Genera and Their Placement in Tribe *Selineae* (*Apiaceae*, *Subfamily apioideae*). *Plant Syst. Evol.* **2004**, *243*, 189–210. [CrossRef]
44. Lohmann, L.G. Untangling the Phylogeny of *Neotropical* Lianas (*Bignonieae*, *Bignoniaceae*). *Am. J. Bot.* **2006**, *93*, 304–318. [CrossRef] [PubMed]
45. Mast, A.R.; Kelso, S.; Richards, A.J.; Lang, D.J.; Feller, D.M.S.; Conti, E. Phylogenetic Relationships in *Primula* L. and Related Genera (*Primulaceae*) Based on Noncoding Chloroplast DNA. *Int. J. Plant Sci.* **2001**, *162*, 1381–1400. [CrossRef]
46. Trift, I.; Källersjö, M.; Anderberg, A.A. The Monophyly of *Primula* (*Primulaceae*) Evaluated by Analysis of Sequences from the Chloroplast Gene RbcL. *Syst. Bot.* **2002**, *27*, 396–407.
47. Schmidt-Lebuhn, A.N.; de Vos, J.M.; Keller, B.; Conti, E. Phylogenetic Analysis of Primula Section Primula Reveals Rampant Non-Monophyly among Morphologically Distinct Species. *Mol. Phylogenet. Evol.* **2012**, *65*, 23–34. [CrossRef]
48. Steinmann, V.W. The Submersion of *Pedilanthus* into *Euphorbia* (*Euphorbiaceae*). *Acta Bot. Mex.* **2003**, *65*, 45–50. [CrossRef]
49. Yang, Y.; Riina, R.; Morawetz, J.J.; Haevermans, T.; Aubriot, X.; Berry, P.E. Molecular Phylogenetics and Classification of Euphorbia Subgenus Chamaesyce (*Euphorbiaceae*). *Taxon* **2012**, *61*, 764–789. [CrossRef]
50. Lammers, T.G. Circumscription and Phylogeny of the Campanulales. *Ann. Mo. Bot. Gard.* **1992**, *79*, 388. [CrossRef]
51. Lanfear, R.; Kokko, H.; Eyre-Walker, A. Population Size and the Rate of Evolution. *Trends Ecol. Evol.* **2014**, *29*, 33–41. [CrossRef]
52. Mehl, T.; Gruenstaeudl, M. Airpg: Automatically Accessing the Inverted Repeats of Archived Plastid Genomes. *BMC Bioinform.* **2021**, *22*, 413. [CrossRef]
53. Turudić, A.; Liber, Z.; Grdiša, M.; Jakše, J.; Varga, F.; Šatović, Z. Towards the Well-Tempered Chloroplast DNA Sequences. *Plants* **2021**, *10*, 1360. [CrossRef]
54. David, C.H. Tukey Exploratory Data Analysis by John W. Tukey. *Biometrics* **1977**, *33*, 311–318. [CrossRef]
55. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef]
56. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638. [CrossRef]
57. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]