


Article

Estimation of Genome Size in the Endemic Species *Reseda pentagyna* and the Locally Rare Species *Reseda lutea* Using comparative Analyses of Flow Cytometry and K-Mer Approaches

Fahad Al-Qurainy ^{1,†}, Abdel-Rhman Z. Gaafar ^{1,*},, Salim Khan ¹, Mohammad Nadeem ¹, Aref M. Alshameri ¹, Mohamed Tarroum ¹, Saleh Alansi ¹, Naser B. Almarri ² and Norah S. Alfarraj ¹

¹ Department of Botany and Microbiology, College of Science bldg5, King Saud University, Riyadh 11451, Saudi Arabia; falqurainy@ksu.edu.sa (F.A.-Q.); skhan2@ksu.edu.sa (S.K.); mnadeem@ksu.edu.sa (M.N.); aalshameri@ksu.edu.sa (A.M.A.); mtarroum@ksu.edu.sa (M.T.); salansi@ksu.edu.sa (S.A.); 438203416@student.ksu.edu.sa (N.S.A.)

² Ministry of Environment, Water and Agriculture, Riyadh 11195, Saudi Arabia; almarri@moa.gov.sa

* Correspondence: agaafar@ksu.edu.sa; Tel.: +966-54-032-9167

† The authors contributed equally to this work.



Citation: Al-Qurainy, F.; Gaafar, A.-R.Z.; Khan, S.; Nadeem, M.; Alshameri, A.M.; Tarroum, M.; Alansi, S.; Almarri, N.B.; Alfarraj, N.S. Estimation of Genome Size in the Endemic Species *Reseda pentagyna* and the Locally Rare Species *Reseda lutea* Using comparative Analyses of Flow Cytometry and K-Mer Approaches. *Plants* **2021**, *10*, 1362. <https://doi.org/10.3390/plants10071362>

Academic Editor: Abdulqader Jighly

Received: 8 June 2021

Accepted: 1 July 2021

Published: 3 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Genome size is one of the fundamental cytogenetic features of a species, which is critical for the design and initiation of any genome sequencing projects and can provide essential insights in studying taxonomy, cytogenetics, phylogenesis, and evolutionary studies. However, this key cytogenetic information is almost lacking in the endemic species *Reseda pentagyna* and the locally rare species *Reseda lutea* in Saudi Arabia. Therefore, genome size was analyzed by propidium iodide PI flow cytometry and compared to k-mer analysis methods. The standard method for genome size measures (flow cytometry) estimated the genome size of *R. lutea* and *R. pentagyna* with nuclei isolation MB01 buffer were found to be 1.91 ± 0.02 and 2.09 ± 0.03 pg/2 °C, respectively, which corresponded approximately to a haploid genome size of 934 and 1.022 Mbp, respectively. For validation, K-mer analysis was performed on both species' Illumina paired-end sequencing data from both species. Five k-mer analysis approaches were examined for biocomputational estimation of genome size: A general formula and four well-known programs (CovEST, Kmergenie, FindGSE, and GenomeScope). The parameter preferences had a significant impact on GenomeScope and Kmergenie estimates. While the general formula estimations did not differ considerably, with an average genome size of 867.7 and 896. Mbp. The differences across flow cytometry and biocomputational predictions may be due to the high repeat content, particularly long repetitive regions in both genomes, 71% and 57%, which interfered with k-mer analysis. GenomeScope allowed quantification of high heterozygosity levels (1.04 and 1.37%) of *R. lutea* and *R. pentagyna* genomes, respectively. Based on our observations, *R. lutea* may have a tetraploid genome or higher. Our results revealed fundamental cytogenetic information for *R. lutea* and *R. pentagyna*, which should be used in future taxonomic studies and whole-genome sequencing.

Keywords: flow cytometry; endemic; ploidy; rare; *Reseda*; genome size; k-mer

1. Introduction

The development of advanced genomic technologies, and the subsequent storm of data from next-generation sequencing (NGS), has been a great asset to genomic research. However, many fundamental issues concerning genomes remain mostly unresolved. One such issue is the largely unexplored amount of DNA (C-value) in most of the higher clades of life. The amount of DNA (C-value) in the haploid gametic nucleus is referred to as genome size [1], which is often quantified in picograms (pg) or megabase pairs (1 pg = 978 Mbp) [2] and is typically broadly constant within an organism [3,4]. Besides

external characteristics, genome size is a key value for research on taxonomy, ecology, and evolution [5,6]. Variations significant enough to differentiate a population into distinct species may still be difficult to discern employing classic morphological or DNA sequence; nevertheless, such variations may become more obvious when genome size is investigated along with other proofs [7,8]. Moreover, a precise calculation of genome size is a prerequisite in the age of high-throughput sequencing technologies for sequencing projects [7], since it influences the budget plan for anticipated sequencing depths and offers an approximate figure for estimating genome assembly completeness.

As a result, there is a great demand for reliable and easy-to-use methods for calculating genome sizes throughout a wide range of eukaryotic taxa [9]. There are two main methods for calculating genome size: Laboratory and computational. The Feulgen microdensitometer and flow cytometry are fairly tested and often used laboratory approaches [10]. Flow cytometry is a low-cost, relatively reliable, and quick laboratory technique for estimating plant genome size. It is an appealing alternative to microspectrophotometry in that it involves the calculation of DNA quantities based on the staining of undamaged nuclei with a fluorochrome that quantitatively adheres to the DNA. Moreover, for the analysis, only a small amount of tissue is needed, which is important in the case of valuable and/or protected specimens [10,11]. These methods, however, rely on living, adequately fixed, or frozen tissues with substantially intact cells, thereby limiting research to lifeforms that can be cultivated in the lab or easily obtained in the field and transferred to the lab [12]. Furthermore, considering that the significant amounts of phenolic compounds can create stoichiometric errors, the flow cytometry approach must be tailored to each plant species.

Meanwhile, with the explosive growth of next-generation sequencing technology, a computational technique arose through k-mer (distinct subsequences of a given length, k, derived from a longer DNA sequence) approaches. A k-mer frequency distribution could be generated by plotting the coverage distribution over all k-mers in a sequence. This k-mer distribution should resemble a Poisson distribution when the created k-mers from genomic sequencing reads possess minuscule amounts of sequence defects (repetitions, sequencing errors, or coverage bias). The distribution peak will be centered on the average sequencing depth for the genome [13]. K-mer based genome size estimates were accurately employed in many genome projects due to their feasibility and rationale [14–17]. Researchers can utilize many available programs to estimate genome size using sequencing data as well as the popular equation, i.e., the quotient of the k-mers total number and the peak frequency distribution. However, the accuracy and efficiency of these strategies have not been thoroughly investigated.

The Resedaceae is a relatively small family with only six genera (i.e., *Reseda*, *Randonia*, *Sesamoides*, *Oligomeris*, *Ochradenus*, and *Caylusea*) and about 85 species [18]. Genus *Reseda* contains approximately 65 species throughout the world, mostly restricted to the Mediterranean basin. Several of its species flourish on soils under arid environments, while others are ruderal weeds and only a few are available in high mountains [19]. Pharmacological studies of various *Reseda* species showed antimicrobial [20], anti-inflammatory [21], and antioxidant [22] activities. In Saudi Arabia, seven species of the genus *Reseda* were recorded, viz. *R. alba*, *R. arabica*, *R. aucheri*, *R. lutea*, *R. muricata*, *R. pentagyna*, and *R. sphenocleoides* [23]. Among these, *R. pentagyna* is endemic and native to Saudi Arabia and has been observed in northeastern region in Tabuk, Wadi Sawawin, and Northern Hijaz Mountain range [24]. *R. pentagyna* is an annual sparsely branched herbaceous plant well adapted to hard sand and low rocky hills with stems erect up to 30 cm, distinguished from the other *Reseda* species via its five to six-toothed capsule [25]. While *R. lutea* L. is locally rare and restricted only to a single gathering in the mountainous region of Abha, Saudi Arabia [26]. *R. lutea* L. is a deep-rooted biennial or perennial herbaceous plant that can grow up to 80 cm high and well adapted to fallow fields, rocky slopes, and roadsides. It is distributed and spread throughout many temperate zones of the world [27]. This study aimed to determine the genome size focusing on the endemic species *R. pentagyna* and the locally rare species

R. lutea both experimentally using flow cytometry and computationally using the k-mer approach through a combination of short-read sequencing with bioinformatics tools.

2. Material and Methods

2.1. Plant Material

Seeds from adult plants of both the endemic species *R. pentagyna* and the rare species *R. lutea* were collected from Abha, Saudi Arabia, for in vitro plant propagation. The identification was confirmed through morphological features coupled with the assistance of Flora of Saudi Arabia [28] and protologue [29], and a voucher specimen (SBSN00015 and SBSN00016) was deposited at the Seed Bank Herbarium, College of Sciences, King Saud University, KSA. The intact seeds were surface-sterilized with 0.3% sodium hypochlorite for 2 to 3 min, then washed 3 to 4 times with double-sterilized water. The seeds were germinated on 2% agar then inoculated on Murashige and Skoog (MS) medium [30].

2.2. Genomic DNA Extraction

A leaf sample from germinated seeds was detached from the medium and directly used for DNA isolation (Figure 1). Total genomic DNA was isolated from *R. lutea* and *R. pentagyna* leaves using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. The NanoDrop2000 spectrophotometer was used to evaluate the purity and amount of DNA (Thermo Fisher Scientific, Waltham, MA, USA). DNA integrity was determined using a 1% (*w/v*) agarose gel electrophoresis. The nuclear ITS region (internal transcribed spacer sequences) was amplified on an AB Veriti 96 well Thermal cycler (Applied Biosystems, Waltham, MA, USA) using PuReTaq Ready-To-Go PCR Beads (GE Healthcare, Little Chalfont, Buckinghamshire, UK). Universal ITS primers were used for amplification and cycle sequencing (ITS1 and ITS4 [31,32]) using the following conditions: Initial denaturation at 94 °C for 5 min, 25 cycles of denaturation for 30 s at 94 °C, annealing at 48 °C for 30 s, extension at 72 °C for 1 min, and a final extension at 72 °C for 7 min. PCR reactions were examined on a 1.2% (*w/v*) agarose gel to confirm the concentration and size of the PCR products. Following standard procedures, Macrogen Inc. (Geumchun-gu, Seoul, South Korea) used a 96-capillary ABI 3730xl DNA analyzer (Applied Biosystems, Foster City, CA, USA) to sequence the amplicons bidirectionally.

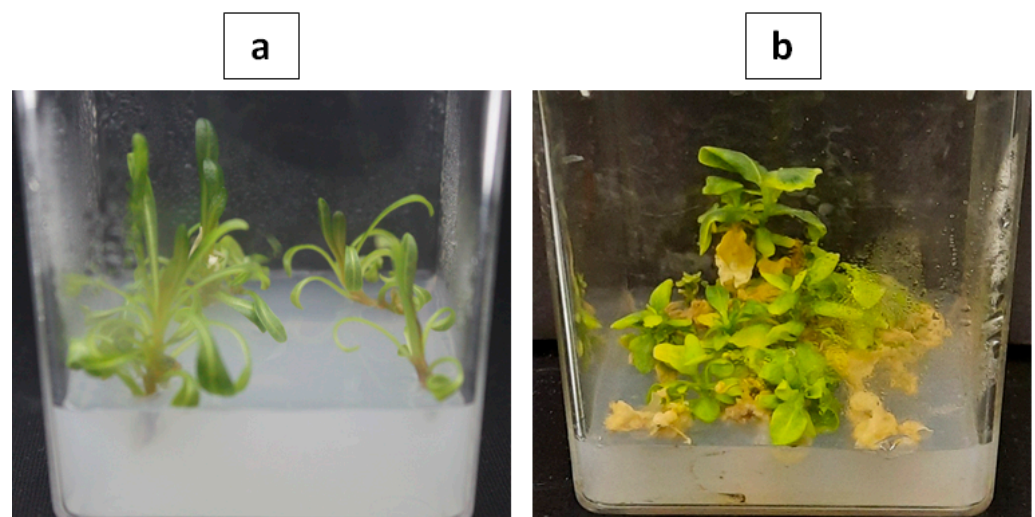


Figure 1. Shoots raised on MS medium (a) *Reseda lutea*; (b) *Reseda pentagyna*.

2.3. Molecular Identification

For molecular identification and phylogenetic assessment, ITS sequences from 50 related *Reseda* species (including representatives from each of the six sections, namely *Resedastrum*, *Phyteuma*, *Neoreseda*, *Luteola*, *Leucoreseda*, and *Glaucoreseda* [18]) were

acquired from GenBank (Figure 2). Sequences of two species from the genus *Stixis* (Resedaceae) were selected and retrieved from GenBank as the outgroup in the phylogenetic analyses (Figure 2). All analyses were implemented in MEGA X [33]. Sequence alignments were performed using Clustal W within the MEGA X windows interface, with manual adjustments. The Neighbor-Joining (NJ) method was utilized for phylogenetic analysis, and the model test was employed to identify the best-fit model for the NJ analysis (Kimura 2-parameter model with a discontinuous Gamma distribution K2 + G). The NJ method was selected for the construction of the phylogenetic tree because it has demonstrated advantages over distance and parsimony approaches to analyze the process of sequence evolution [34]. To obtain statistical support for every internal and external branch, a bootstrap test with 2000 replication was run concurrently for all analyses.

2.4. Flow Cytometric Genome Size

The young leaves from multiple shoots raised on MS media were used for the extraction of nuclei. Dr. Jaroslav Dolezel (Laboratory of Molecular Cytogenetics and Cytometry, Institute of Experimental Botany, Sokolovská 6, Olomouc, Czech Republic) kindly offered the seeds of external reference *Solanum lycopersicum* cv. Stupické (2C = 1.96 pg) [35]. MB01 buffer [36] was used for the estimation of 2C DNA content of *Reseda lutea* and *Reseda pentagyna* (2.5 mM Na₂EDTA; 20 mM MOPS; 0.2% (v/v) Triton X-100; 80 mM KCl; 0.7 mM Spermine tetrahydrochloride; 20 mM NaCl; pH 7.4). In addition, antioxidants including 1% PVP and 0.5% β-mercaptoethanol were freshly prepared and added for extraction of pure nuclei.

All experiment steps of nuclei extraction were performed on ice (4 °C). The young leaves (30 mg) were chopped with a sharp razor blade into 0.3–0.6 mm size in a petri dish containing ice-cold 500 µL MB01 nuclei isolation buffer. The suspension was mixed by pipetting and filtered through a 20 µm double nylon mesh. After filtration, the nuclei suspension was stained for 10 min with 50 µg/mL of PI (Propidium iodide, Sigma, St. Louis, MO, USA) under dark refrigeration, and the samples were stored on ice prior to analysis.

The fluorescence of a minimum of 5000 propidium iodide-stained nuclei was estimated using a flow cytometer Muse cell analyzer (Merck Millipore, Burlington, MA, USA). The flow rate of the capillary was set at 0.12 µL/s, which is very low. Propidium iodide was measured at 585 nm to read the 2C nuclei DNA content of the sample. The obtained histograms were computerized by Muse cell analyzer software package (Muse 1.8 analyses, Burlington, MA, USA). The sample 2C DNA content was calculated according to the formula [37]:

$$2C \text{ DNA content of sample} = \frac{(\text{Fluorescence mean intensity of sample})}{(\text{Fluorescence mean intensity of standard})} \times 2C \text{ DNA content of standard} \quad (1)$$

The number of base pairs per haploid genome was determined using the formula 1 pg DNA = 978 Mbp [2,38]. Three replicate measurements were taken for each plant species independently. The fluorescence histograms were resolved into G0/G1 (2C), S, and G2/M (4C) cell-cycle compartments. The fluorescence mean intensity was taken for the calculation of the 2C DNA content of *Reseda* species. To improve accuracy, the genome size was determined for each sample as the mean of two technical and three biological replicates, enabling the standard error to be calculated.

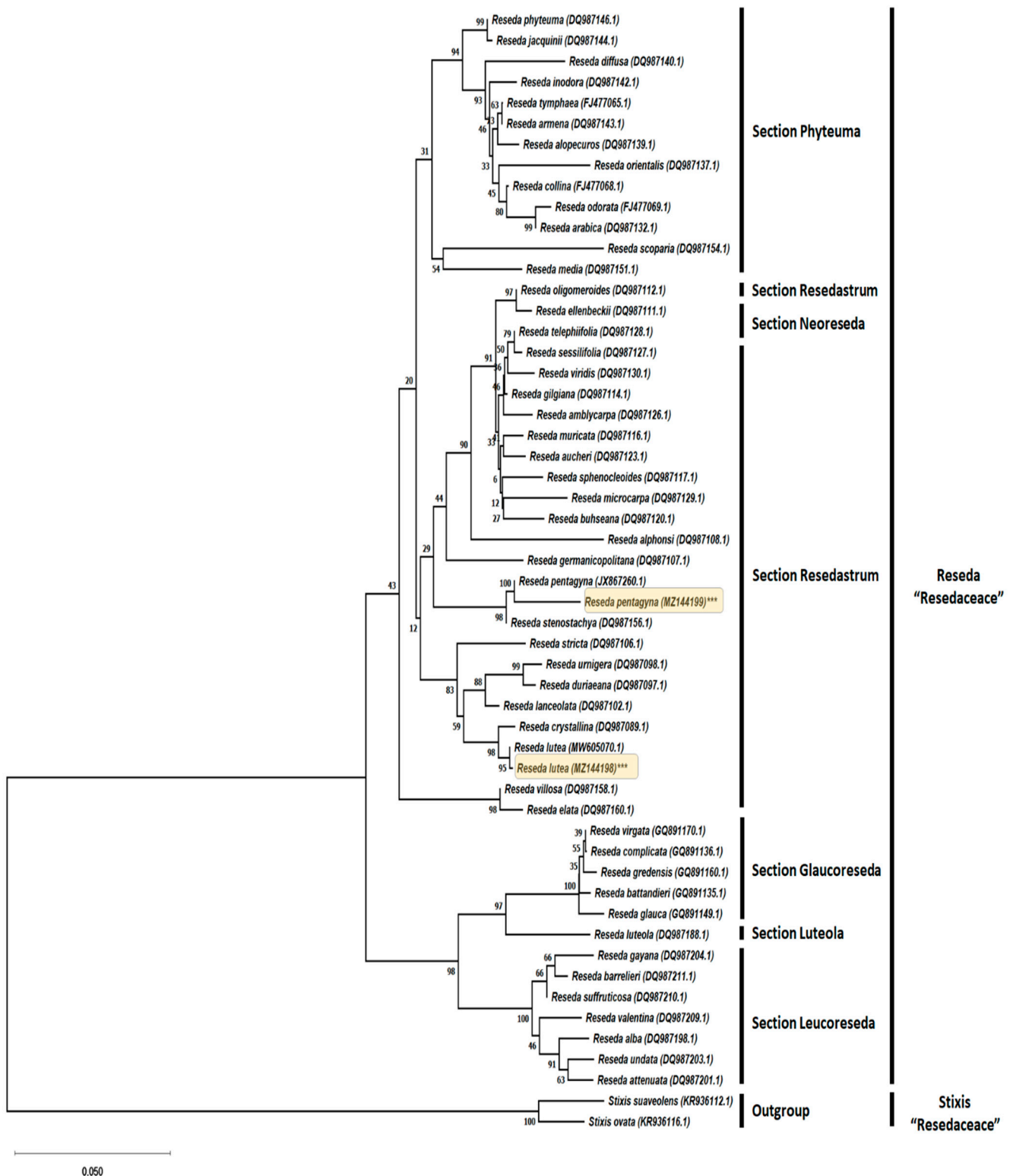


Figure 2. Molecular identification of *Reseda lutea* and *Reseda pentagyna* (highlighted) was conducted via evolutionary analyses in MEGA X and inferred from the analysis of internal transcribed spacer (ITS) sequence using the Neighbor-Joining algorithm. Next to the branches is the percentage of replicate trees around which the linked taxa grouped with each other in the bootstrap test (2000 replicates). The Kimura 2-parameter model was used to calculate the evolutionary distances (NCBI accession numbers between brackets).

2.5. Whole-Genome Sequencing and Filtering Contaminated Reads

Macrogen Inc. prepared the DNA libraries for genome sequencing (S. Korea). Using the TruSeq Nano DNA kit, a paired-end 350 bp insert size library was created for the two species (Illumina, San Diego, California, USA). The libraries were then sequenced using 2×151 bp paired-end sequencing on the Illumina NovaSeq6000 platform using standard Illumina operating protocol, yielding a minimum of 90 Gb of raw data. The run's primary data processing was completed with the manufacturer's program Real-Time Analysis (RTA 1.18.66.3), followed by the construction of FASTQ sequence files with the Illumina tool bcl2fastq. The raw sequencing reads were deposited in the GenBank database under the BioProject accession PRJNA733338. FastQC v0.11.9 [39] was used to visually examine the raw read quality. Trimmomatic (v0.38) [40] was used to delete the remaining adapter sequences, leading and trailing nucleotides with a Phred score of less than 25, and reads less than 50 bp. SOAPec v2.03 [41] was used to fix the errors in filtered reads. FastUniq v1.1 [42] was used to remove duplicated read pairs. All reads were filtered of potential contaminants by mapping via the BBDuk module (BBMap v38.9 [43]) against a contamination database that included chloroplast, mitochondrial, bacterial, and viral sequences, etc., detected with FastQ Screen [44] keeping only unmapped reads and subsequently assessed again using FastQC.

2.6. K-Mer Based Genome Size

Even though the genome size can be calculated by tallying the k-mer frequency of the read data, the k-mer must be high enough to differentiate most of the genome. The optimal k-mer length for genome size estimation has not been extensively tested. The k-mer value varies amongst investigations, whereas values between 17 and 35 are prominent [45,46]. At least 17 are commonly employed in most eukaryotic genomes to prevent palindromic sequences and the effect of excessively repetitive DNA sequences. For analysis, first the frequency distribution of three k-mers (i.e., 21, 31, and 41) was generated using Jellyfish v2.3.0 [47]. Second, four k-mer analysis-based methods were evaluated for computational genome size estimation, including the most recent dedicated tools (Kmergenie v1.7 [48], GenomeScope v1 [49], FindGSE v1.94 [50], and CovEST-repeat [51]) and the commonly used formula for the calculations of genome size sourced from the equation ($M = N \times (L - K + 1)/L$) proposed by the M.S. Waterman group, where (M) the reads k-mer frequency peak is associated with (N) the actual sequencing depth, (K) kmer length, and (L) read length [13,52,53]. Third, the ploidy structure was estimated with Smudgeplot v0.2.3 [54]. Finally, GenomeScope v1 was run using k-mer length ($k = 21$) and analyzed the histograms to estimate the complexity of the genome (heterozygosity and repeats) with maximal k-mer coverage = -1 .

3. Result and Discussions

3.1. Molecular Identification

Internal transcribed spacer ITS sequences of nuclear ribosomal DNA have received a lot of attention over the last two decades, not only because of their effectiveness in performing plant phylogeny at a lower taxonomic level, but also because they are regarded as far more reliable markers available for plant DNA barcoding. Due to the highly intriguing morphological similarities reported across *Reseda* species [28,29], molecular identification and phylogenetic analysis with ITS were implemented to determine the species designation of *R. lutea* and *R. pentagyna*.

To validate the morphology-based taxonomic identification of *R. pentagyna* and *R. lutea*, the ITS region was sequenced and aligned to 50 *Reseda* species with ITS sequences currently available at NCBI (including the ones for both *R. lutea* and *R. pentagyna*). The combined length of ITS region for the two plants comprised 699 and 707 nucleotides, respectively. A BLAST screening of *R. pentagyna*'s ITS query sequence revealed the highest sequence identity and similarity to previously published *R. pentagyna* ITS sequences JX867260.1 97.95% and similarly *R. lutea* 99.86% for itself KR936125.1.

The Neighbor-Joining algorithm was used to infer the evolutionary phylogram tree with the lowest BIC (Bayesian Information Criterion) score of 9855.053 based on the Kimura 2-parameter model to estimate a matrix of pairwise distances. The evolutionary rate differences between sites were modeled using a discrete Gamma distribution (5 categories). The tree is depicted to scale, and branch lengths are calculated by counting the number of substitutions for each site. The tree was rooted with the help of *Stixis suaveolens* (KR936112.1) and *Stixis ovata* (KR936116.1) as an outgroup. Bootstrap supports (%) with a value greater than 50% are displayed above branches.

The Neighbor-Joining tree derived from the analysis of ITS sequences is in line with previous phylogenetic analyses and revealed grouping of *Reseda* species consistent with established taxonomic sections of the genus, *R. pentagyna* showed proximity with *R. stenotachya* (98% bootstrap support), while *R. lutea* showed proximity with *R. crystallina* (99% bootstrap support) nested within the clade of section Resedastrum (Figure 2). The research concluded that *Reseda* species were grouped and consistent with preexisting taxonomic sections [18]. As a result, our ITS analysis validated the taxonomic identification and classification of the examined plants based on morphology.

3.2. C-Value Determination via Flow cytometry

Due to the development of flow cytometry, the study of genome size and its significance has dramatically increased in recent years not just as a taxonomic marker, but also for assessing how it corresponds to environmental, ecological, and phenotypic variables [55–58]. Furthermore, before determining the nucleotide sequence of a plant's DNA, it is necessary to understand how large the genome is [59]. According to a large-scale analysis of plant genome sizes, large genomes are less resistant to environmental pressures like drought or pollution, and are less capable of adjusting, making them more vulnerable to extinction [60,61]. Consequently, the genome size evolution heads toward small genomes [59]. Therefore, knowledge of the genome size of the two species of *Reseda* under study could be used for the prediction of the threat of extinction particularly the rare species *R. lutea* [60].

Preliminary testing revealed the success of flow cytometry analyses with both *Reseda* species forming peaks in the histograms. The 2C peaks in the histograms for fresh plant materials were suitable for genome size estimation (Figure 3). The nuclear DNA content of the two species of *Reseda* was evaluated by flow cytometry using tomato (2C = 1.96 pg) as an external reference standard, which was later determined to be the most appropriate standard for *Reseda* samples due to their proximate DNA content. The genome size for *R. lutea* and *R. pentagyna* showed a narrow range and was estimated to be 1.91 ± 0.02 and 2.09 ± 0.03 , respectively (Table 1). Our estimations for *R. lutea* and *R. pentagyna* constitute one of the highest values so far for this genus *Reseda* (0.92–2.86 pg/2C). For *R. lutea*, whose genome size had previously been assessed, there was a clear agreement with earlier findings (Table 2). The slight difference in DNA content could occur due to the type of laser lamp equipped in the flow cytometer [62]. According to Soltis et al. [63] classification, both *Reseda* species genomes belong to the category of plants with a smaller genome. The genome of *R. pentagyna* is around the same size as that of *R. lutea*, an octoploid species [18]. Furthermore, its genome is nearly twice as large as that of *R. suffruticosa*, which possesses a tetraploid genome [18,64].

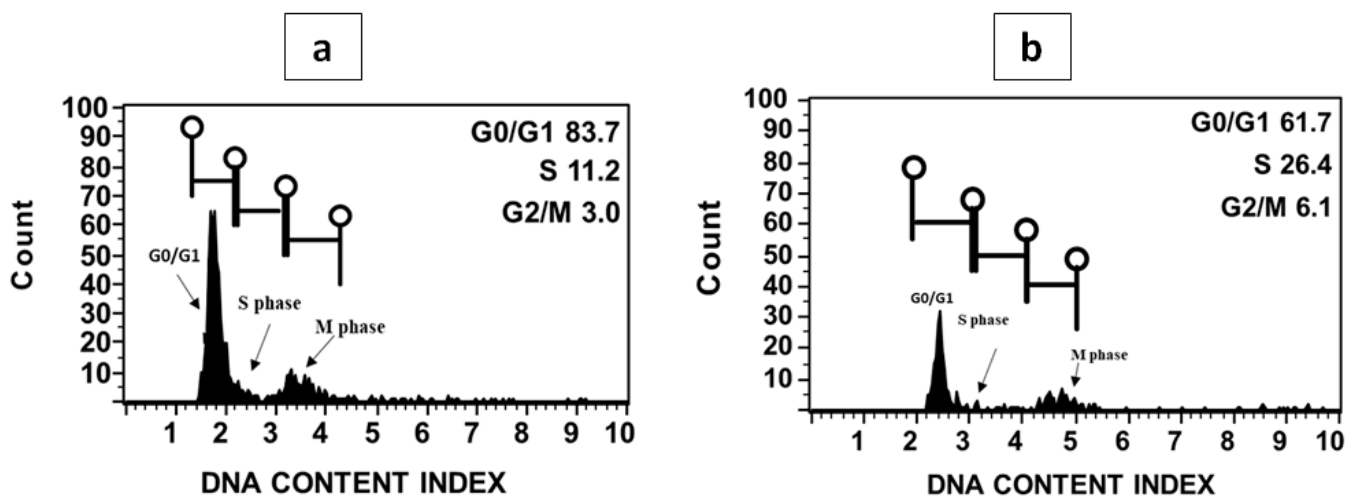


Figure 3. Histogram of fluorescence intensity for genome size assessments in *R. lutea* (a) and *R. pentagyna* (b) nuclei stained with propidium iodide prepared from shoot tissues. The two major phases of the cell cycle (interphase G0, G1, S, G2 and the mitotic phase M).

Table 1. C-value comparison (flowcytometry vs. k-mer) genome size and G0/G1 phases of the cell cycle in *Reseda lutea* and *Reseda pentagyna*. Genome size (mean \pm standard deviation). The number of individuals analyzed for genome size (n).

	n	G0/G1 (%)	2C DNA Content (pg)	1C DNA Content (pg)	1C DNA Content (Mbp)	2C K-Mer (pg)	1C K-Mer (pg)	1C K-Mer (Mbp)
<i>Reseda lutea</i>	3	86.53 \pm 2.20	1.91 \pm 0.02	0.955 \pm 0.01	973.11	1.78	0.89	867.7
<i>Reseda pentagyna</i>	3	80.96 \pm 0.83	2.09 \pm 0.029	1.045 \pm 0.015	1026.9	1.84	0.92	896.3

* The external reference standard [*Solanum lycopersicum* (2C = 1.96 pg)] * c-values with the two methods are significantly different at $p < 0.01$ (one-sample t -test).

Table 2. Cytogenetical characteristics of *Reseda* species [64,65].

	Species	2C (pg)	Chromosome Number	Ploidy Level	Basic Chromosome Number	Section
1	<i>R. lutea</i>	2.06	24, 48	4-8	6	Resedastrum
2	<i>R. stricta</i>	2.86	24	4	6	
3	<i>R. lanceolata</i>	1.70	24	4	6	
4	<i>R. odorata</i>	0.96	12	2	6	Phyteuma
5	<i>R. phyteuma</i>	1.34	24	4	6	
6	<i>R. media</i>	2.09	12	2	6	Leucoreseda
7	<i>R. undata</i>	1.22	20	4	5	
8	<i>R. barrelieri</i>	1.68	20	4	5	
9	<i>R. suffruticosa</i>	0.92	20	4	5	Luteola
10	<i>R. alba</i>	1.45	40	8	5	
11	<i>R. luteola</i>	1.75	24	4	6	
12	<i>R. glauca</i>	2.11	28	4	7	Glaucoreseda
13	<i>R. complicata</i>	1.71	28	4	7	
14	<i>R. virgata</i>	1.44	28	4	7	
15	<i>R. gredensis</i>	2.63	28	4	7	

3.3. Whole-Genome Sequencing

The development of improved sequencing technology capable of producing considerable amounts of sequence data at a low cost, combined with enhanced assembling procedures, has expanded both model and non-model plants genome sequencing [66]. The paired-end 350 bp insert size libraries (Figure 4) of *R. lutea* and *R. pentagyna* were sequenced using the HiSeq 2500 Illumina sequencing platform, which produced 358.2 and 352.4 million pairs of 151bp reads, accounting for a total of 108.2 Gb and 106.4 Gb of sequence, respectively. Based on the flow cytometry estimates of genome size, the sequence data represented more than 100× coverage of both genomes (Table 3). Tools for estimating genome size employing k-mer distributions perform much better whenever the average coverage is higher than 10× [67]. Quality filtering (removing bases with a Phred score of less than 25 and reads shorter than 50 bp) did not significantly decrease the dataset. Approximately 0.6–1.5% of the reads identified by FastQ Screen (Figure 5) as contaminants (chloroplast, mitochondrial, bacterial, and viral sequences, etc.), which in turn were used to map the clean reads with bbdutk2, leaving between 637.9 Mbp and 632.3 Mbp unmapped reads for further processing (Table 3 and Figure 6). After the quality filtering was established, the raw data mean read length was 148bp.

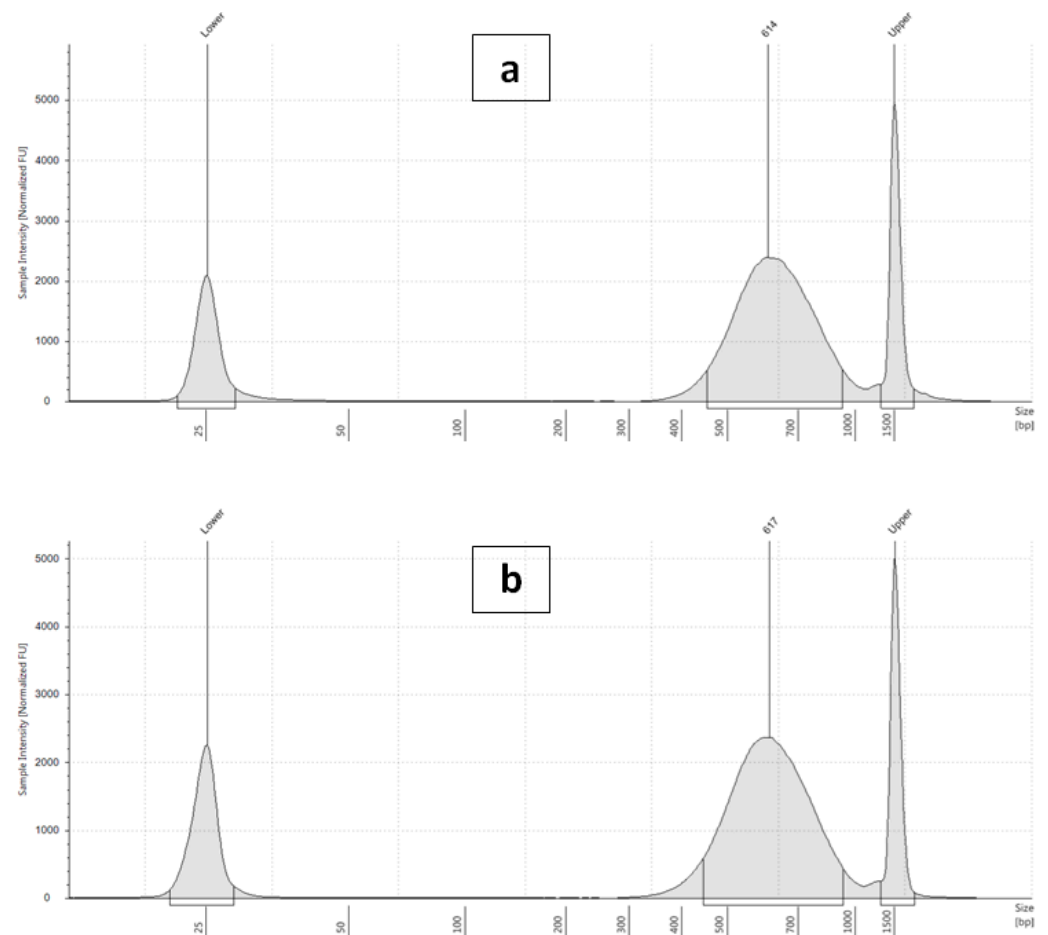


Figure 4. Quality check electropherogram of post-library construction on the Agilent 2200 TapeStation (a) *Reseda lutea* and (b) *Reseda pentagyna*.

Table 3. Summary statistics for *Reseda lutea* and *Reseda pentagyna* genome sequences.

Sample ID	Fragment Length (bp)	Read Length (bp)	Total Reads	Clean Unmapped Reads ^a	GC(%)	AT(%)	Q20(%)	Q30(%)
<i>Reseda lutea</i>	614	2 × 151	716,375,240	637,861,144	45.01	54.99	95.83	90.59
<i>Reseda pentagyna</i>	617	2 × 151	704,839,182	632,346,592	52.83	47.17	97.18	92.59

^a Clean reads: The number of reads that have survived after quality trimming and contamination removal.

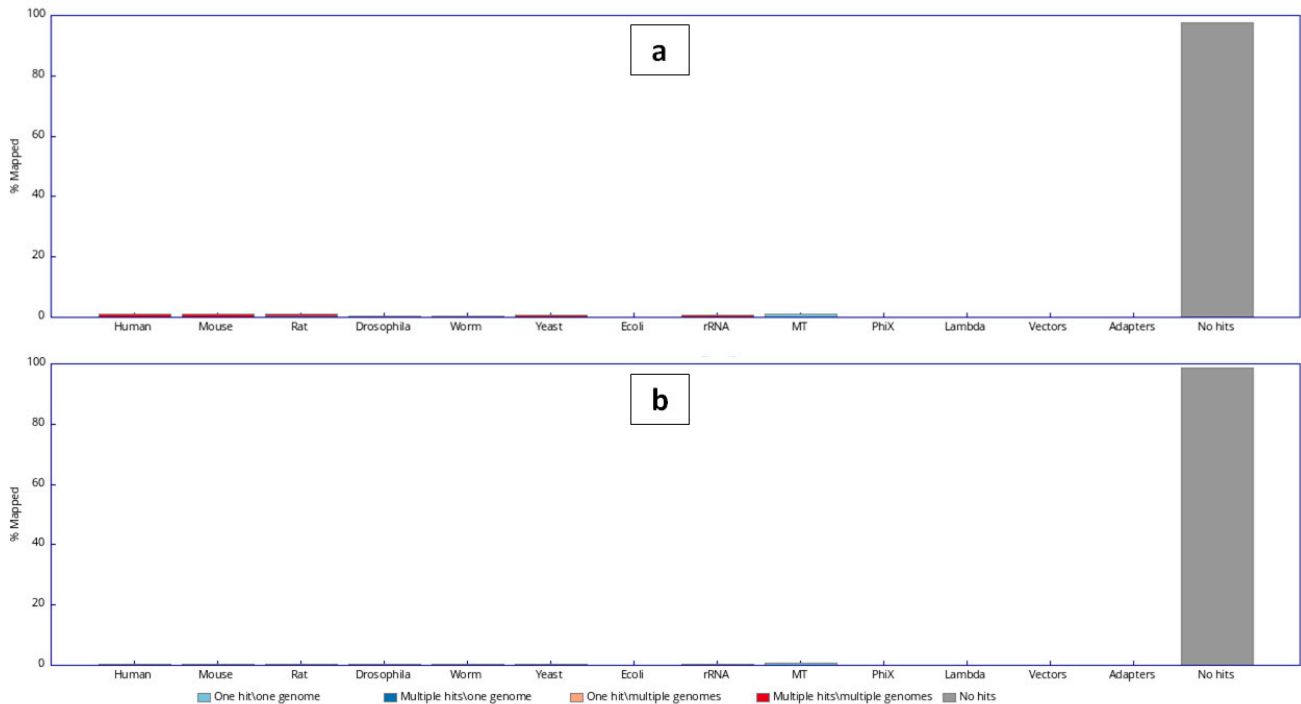


Figure 5. The plot shows the frequency distribution of the contaminating taxa for (a) *Reseda lutea* and (b) *Reseda pentagyna*.

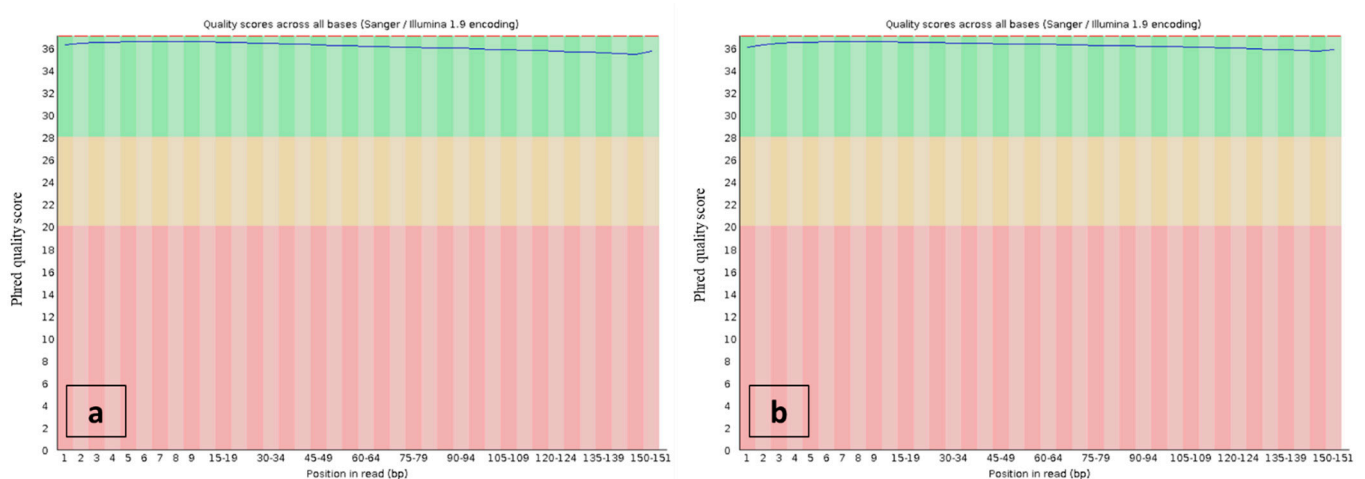


Figure 6. Phred quality scores throughout the read bases—dataset ready for genome size estimation—*Reseda lutea* (a) and *Reseda pentagyna* (b).

3.4. K-Mer Based Genome Size and Complexity

Accurate genome size measurement is crucial for genome research projects [1], and it provides data for analyzing variation in genome size over a wide taxonomic group [68]. Nevertheless, calculating genome size effectively with flow cytometry demands the elimi-

nation of potential erroneous sources [12,69,70]. Flow cytometry analysis may exaggerate the measured values due to the impact of various plant metabolites on stain binding. Consequently, k-mer analysis was carried out to validate the flow cytometry findings. Although estimates based on k-mer analysis may vary depending on the program's parameter choices, the quality of the sequencing data may also hold a role. Hence, four methods were investigated for computational genome size prediction using k-mer analysis, including the most notable trusted programs (CovEST-repeat, kmergenie, GenomeScope, and FindGSE) and the widely used equation for genome size calculations sourced from the formulas proposed by M.S. Waterman group. The GenomeScope authors suggested k-mer 21 as an acceptable compromise between both computation accuracy and speed [49], while k-mers ranging from 17–27 have been employed in other research [45,50,71]. In this study, all k-mer evaluations were executed with k values ranging from 21–41 to ensure that the k length had no effect on the estimations. The impacts of k-mer size (21-mer, 31-mer, and 41-mer) and raw vs. quality processed data were explored for each program (Table 4). The differences between raw and quality processed datasets were minor and skewed in favor of processed data.

Table 4. K-mer estimations of genome size (Mbp) utilizing raw (R) and quality processed (P) sequencing data for *Reseda lutea* and *Reseda pentagyna*.

<i>Reseda lutea</i>							
Genome Estimation Software	K21		K31		K41		Average Processed data (SD)
	R	P	R	P	R	P	
General Formula	845	851	860	868	876	884	867.7 (16.5)
FindGSE	864	876	972	988	1077	1078	980.7 (101.2)
Covest-Repeat	826	772	885	958	1123	1209	979.67 (219.3)
Kmergenie	391	401	471	483	542	559	447.7 (132.6)
GenomeScope V1	584	591	652	665	788	796	684 (103.8)
<i>Reseda pentagyna</i>							
Genome Estimation Software	K21		K31		K41		Average Processed data (SD)
	R	P	R	P	R	P	
General Formula	871	880	874	882	931	927	896.3 (26.6)
FindGSE	768	781	825	848	935	971	866.7 (96.4)
Covest-Repeat	817	825	1010	1067	1249	1318	1070 (246.51)
Kmergenie	484	486	582	591	611	614	552.3 (66.4)
GenomeScope V1	515	524	602	619	723	748	630.3 (112.4)

(K21) (K31) (K41) k-mer sizes; (SD) Standard Deviation.

According to our findings, the behavior of kmergenie and GenomeSope performance was drastically affected by increasing k-mer. The GenomeSope genome size estimates in processed data varied from 591 Mbp to 796 Mbp in *R. lutea* and from 524 Mbp to 748 Mbp in *R. pentagyna*. A closer examination of the kmergenie results revealed that the predicted genome was roughly half the output expected for both species' haploid genomes, resulting in an underestimated genome size. This was also demonstrated in investigations with cane toad [72], vanilla [73], and Pacific oyster [49,74], where k-mer-based GenomeScope estimations of genome sizes were barely half of those derived by flow cytometry and far smaller than those achieved after genome assembly. The discrepancy demonstrates that these strategies might be unreliable in some instances. The genome size estimates from the other k-mer methods were generally slightly low compared to flow cytometry estimates, but different from CovEST "repeat" estimates, which were higher on average than the size

suggested by the flow cytometry measurements. Such a basic pattern was also detected while matching whole-genome assemblies to flow cytometry and Feulgen staining [75].

GSE predicted *R. lutea* genome size of average 980.7 ± 101.2 Mbp while 886.7 ± 96.4 Mbp in *R. pentagyna*. For both genomes with the General Formula prediction, the effects of using different kmer sizes were small (<0.016 Gbps). With this Formula, the genome size estimates for *R. lutea* and *R. pentagyna* of 907 ± 16.5 Mbp and 896.3 ± 26.6 Mbp. In general, the haploid genome size estimation of *R. lutea* based on k-mer distributions of the Illumina sequence reads ranged from 447.7 Mbp (kmergenie), over 680 Mbp (GenomeScope), over 860 Mbp (General Formula), to over 950 Mbp (FindGSE) while *R. pentagyna* ranged from 552.3 Mbp (kmergenie), over 630.3 Mbp (GenomeScope) to around 900 Mbp (FindGSE, General Formula) (Table 4).

The k-mers depth distribution histograms (Figure 7) revealed a unique bimodal profile in both species with a high peak around $40\times$ coverage and a shorter peak around $80\times$. This could be evidence of a highly heterogeneous genome [49]. Additionally, GenomeScope estimated that all genomes consisted of high repetitive sequences. Values of lower k-mers yielded much lower genome size estimates than suggested by flow cytometry, while larger k values produced estimates that were more consistent. The C-value determined by the General Formula k-mer analysis average value for *R. lutea* was 1.78 pg/2C, which is 0.13 pg lower than the C-value determined by flow cytometry. The proportion of repetitive sequences was determined to be approximately 71.26% based on the distribution of k-mers, while heterozygosity was approximately 1.04% (Table 5). The proportion of repetitive sequences and heterozygosity in *R. pentagyna* were approximately 56.77% and 1.37%, respectively. The C-value based on k-mer analysis was 1.84 pg/2C, which is 0.25 pg lower than that predicted from flow cytometry (Table 1). Similar inconsistencies have been documented for *Arabidopsis thaliana*, as well as European eels, and were attributed to chemical compounds interference in stoichiometric DNA content estimations in flow cytometry analysis [50,76].

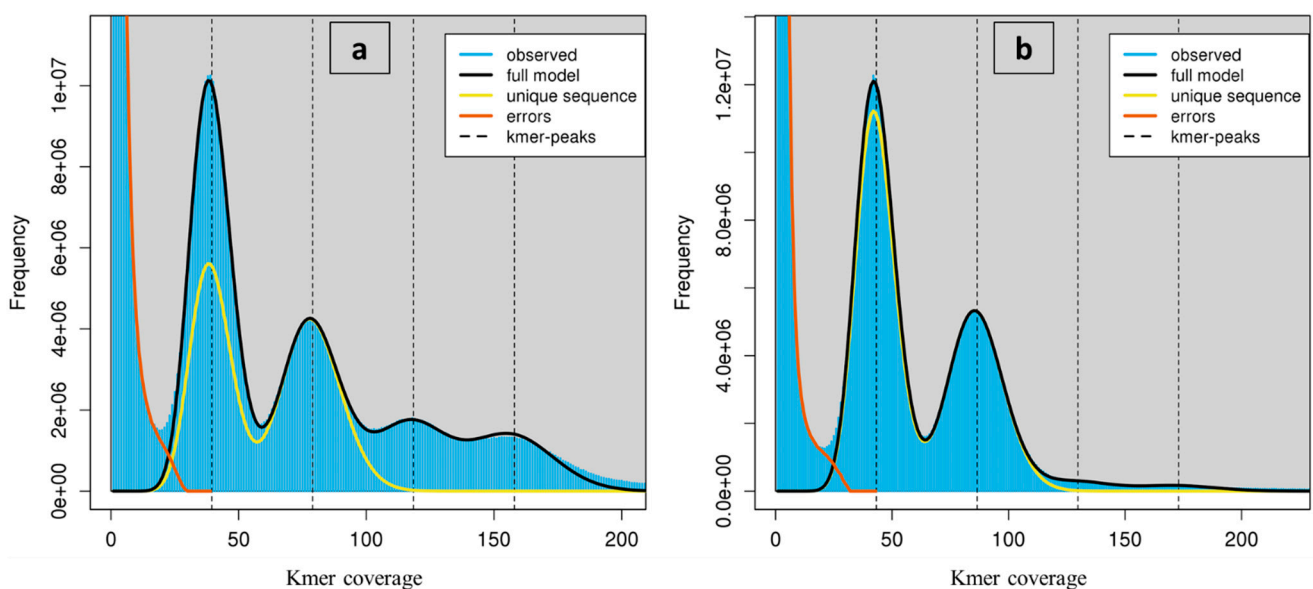


Figure 7. K-mer profile ($k = 41$) spectrum analysis to estimate genome size in *Reseda lutea* (a) and *Reseda pentagyna* (b) generated from sequence data using GenomeScope V1. The high peak at quite low depths, induced by sequencing errors, has been trimmed to empower visualization.

Table 5. Genome properties of *Reseda lutea* and *Reseda pentagyna*.

Genome size Property *	<i>Reseda lutea</i>		<i>Reseda pentagyna</i>	
	min	max	min	max
Homozygous (%)	98.96	98.96	98.63	98.63
Heterozygous (%)	1.04	1.04	1.37	1.37
Genome Haploid Length (bp)	789,888,133	796,236,693	747,545,978	747,661,754
Genome Repeat Length (bp)	562,888,521	566,147,654	424,413,611	424,480,553
Genome Unique Length (bp)	226,999,612	230,089,039	323,132,368	323,181,201
Model Fit (%)	91.35	97.38	95.13	98.3
Read Error Rate (%)	0.06	0.06	0.12	0.12
Repeats (%)	71.26	71.1	56.77	56.77

* Estimated from processed reads by GenomeScope v1 with k = 41.

However, the observed slight difference in genome size estimated for *R. lutea* and *R. pentagyna* when determined through using k-mer and flow cytometry methods could be attributable to the comparatively low sequencing depth as well as the relatively significant proportion of complex or long repetitive elements (>short reads) in these species' nuclear genomes (Table 5) that were not recovered in the sequencing [77]. According to Kidwell [78], there is a close association between repetitive DNA sequences and genome size, and the link was demonstrated by Li et al. [79]. Once they account for a high fraction of the genome, repetitive elements are known to restrict genome size estimates downwards [80].

In maize, repeats account for approximately 80% [81] of the genome, with a sophisticated structure that complicates whole-genome sequencing [82]. These constraints could be quickly overcome with further participation of deep sequencing from third-generation sequencing technology. Additionally, the substantial genome size estimated for *R. lutea* and *R. pentagyna* (≈ 1 Gbp) indicates that constructing a high-quality (i.e., chromosomal level) genome will most likely require a combination of short and long reads (i.e., ONT, PacBio). Long reads with lengths of ~ 10 – 20 kbp [83] can allow clarification of repetitive genomic zones, while short reads, in turn, increase assembly accuracy since their error rate is relatively lower than long reads ones [84,85].

The kmer length had a significant impact on predicted genome size in both species (p -value > 0.01 —one-way ANOVA).

3.5. Ploidy Level Estimation

Detailed bibliographic research on the documented basic chromosome number and ploidy levels of the examined taxa was performed to determine the DNA ploidy level. In terms of chromosomal numbers and ploidy level (mostly from the following online databases and bibliography: Plant DNA C-values Database [86], Chromosome Counts Database (CCDB) [87], and Index of Plant Chromosome Numbers [88]).

In these studies, the basic chromosome number was proposed to be ($\times = 6$) within the Resedastrum section (Table 2) with two ploidy levels (terta-, octoploid) [64,89], and species possessing chromosome counts $n = 24$ or more were proposed to have evolved from interspecific hybridization and the generation of reproductive plants through hybrid genome doubling [65]. Previously reported chromosome counts for *R. lutea* have been inconsistent [90] with most reports determining its chromosome number to be $2n = 48$ [25,64] whereas few studies identified the chromosome numbers to be 24 [25]. Considering the documented chromosomal counts in *R. lutea* and its unavailability in *R. pentagyna* and depending on the comparable C value among both species, we hypothesize that these species possess the same number of chromosomes, 48.

Moreover, the wide range of DNA content between species (0.92–2.86 pg/2C) in the genus *Reseda* usually supports changes in ploidy, hence Illumina reads were used to assess the ploidy level via Smudgeplot, which uses the ratio of heterozygous k-mer pairs to

estimate ploidy. Analysis of *R. lutea* sequence data provided hints that a polyploid genome, from analysis with a k-mer size of 21 with the most abundant k-mer pairs, is the hexaploid heterozygous (AAAAAB) form of *R. lutea* (Figure 8). To our knowledge, this is the first report of a hexaploid form of *R. lutea*. The probability of *R. lutea* possessing a polyploid genome has been implied based on the genome size expansion and the increase in the basic chromosome numbers.

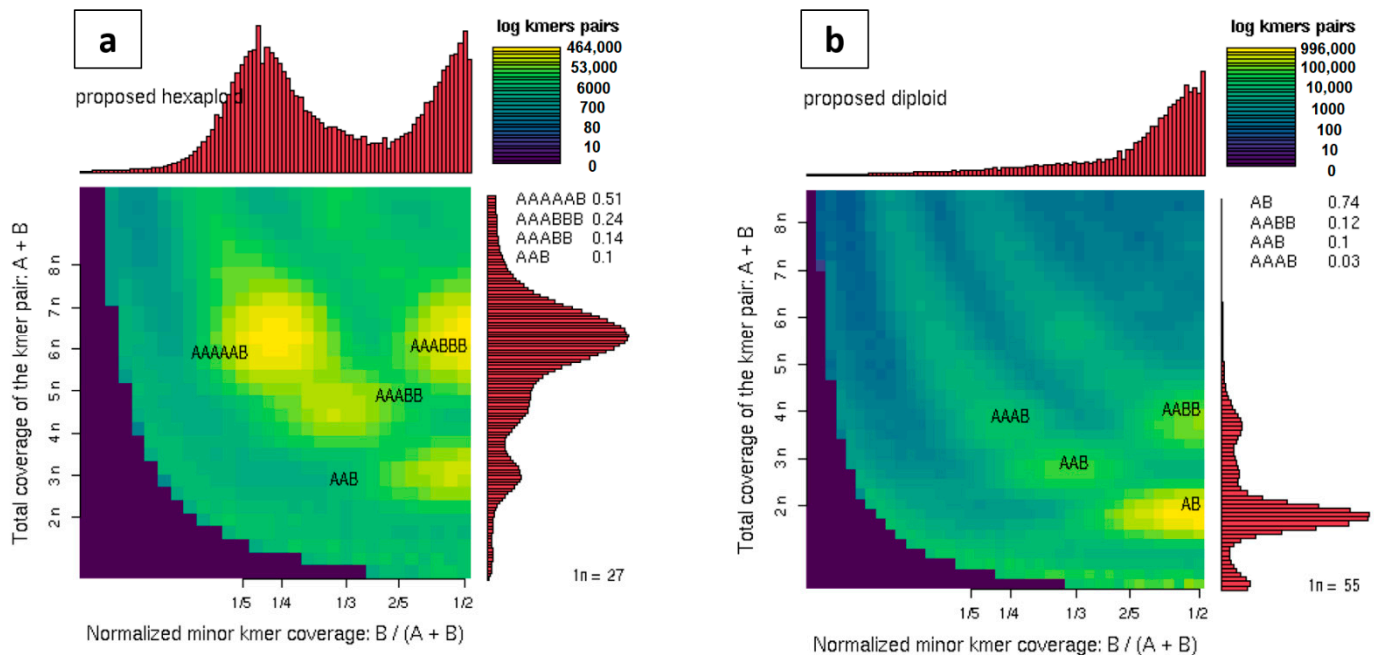


Figure 8. Two-dimensional heat maps were constructed to depict the prediction of ploidy from clean reads using Smudgeplot ($k = 21$). (a) *Reseda lutea* and (b) *Reseda pentagyna*. The color intensity corresponds to the approximate amount of k-mers per bin, ranging from purple (weak) to yellow (strong). Estimated ploidies are shown in the upper left corner of each graph, with the likelihood of various ploidies shown on the right.

Meanwhile, *R. pentagyna* Smudgeplot analysis supported a diploid heterozygous genome (AB) and not polyploidy, which may be the result of the occurrence of a strict uncommon autopolyploid phenomenon that has been revealed in some species [91] and the analysis tool could not interpret. Smudgeplot is designed to predict high heterozygous species and therefore fails to interpret a totally homozygous polyploid genome [54]. However, more cytological studies should be carried out to confirm the chromosome number and verify the ploidy type. Furthermore, because there is a good association between DNA content and ploidy within a species, population-size studies using flow cytometry could be undertaken in the future to differentiate ploidy levels within a species.

4. Conclusions

The significance of the genome size trait is self-evident, as it not only determines plant community configurations at the ecological level, but also impacts plant genome evolution. In this study, the first published flow cytometry estimate for *R. pentagyna* and a confirmation of the previously reported estimate for *R. lutea* were presented alongside the validation and comparison against the estimates via the exploitation of short-read sequence data k-mer analysis. However, some k-mer-based tools demonstrated consistency with flow cytometry estimates. Unfortunately, k-mer analysis remains problematic since its estimates fluctuate based on the tool parameter choices as well as coverage and quality of reads. When fresh material and enough resources are available, flow cytometry should be the preferable method for determining genome size, and kmer should be used solely to provide an approximate estimate. Furthermore, the substantial proportion of repeated

elements identified in both species could imply that the expanded genome resulted from repetitive element amplification along with polyploidization. Based on our results and the rise in chromosome number, we hypothesize that *R. lutea* has a tetraploid genome or higher. More research is needed, however, to validate the ploidy type. The information acquired from this study should provide a basis for future phylogenetic and evolutionary studies, as well as the initiation of genome sequencing projects at the chromosome level.

Author Contributions: Conceptualization, F.A.-Q. and A.-R.Z.G.; methodology, A.-R.Z.G. and S.K.; software, A.-R.Z.G.; validation, A.-R.Z.G., S.K. and A.M.A.; formal analysis, A.-R.Z.G.; investigation, A.-R.Z.G. and S.K.; resources, S.A., M.T. and N.S.A.; data curation, A.-R.Z.G.; writing—original draft preparation, A.-R.Z.G. and S.K.; writing—review and editing, A.-R.Z.G.; visualization, A.M.A.; supervision, F.A.-Q.; project administration, M.N. and N.B.A.; funding acquisition, F.A.-Q. All authors have read and agreed to the published version of the manuscript.

Funding: This Project was funded by the National Plan for Science, Technology, and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (BIO724).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data which support the results of this study are available in GenBank at the NCBI (<https://www.ncbi.nlm.nih.gov>, accessed on 3 July 2021) under the BioProject accession PRJNA733338. NCBI accession numbers for all species in the molecular identification analysis are available in Figure 2.

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

References

- Doležel, J.; Bartos, J. Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Bot.* **2005**, *95*, 99–110. [[CrossRef](#)]
- Doležel, J.; Bartos, J.; Voglmayr, H.; Greilhuber, J. Nuclear DNA content and genome size of trout and human. *Cytom. Part A* **2003**, *51*, 127–128.
- Greilhuber, J.; Doležel, J.; Lysak, M.A.; Bennett, M.D. The origin, evolution and proposed stabilization of the terms ‘genome size’ and ‘C-value’ to describe nuclear DNA contents. *Ann. Bot.* **2005**, *95*, 255–260. [[CrossRef](#)] [[PubMed](#)]
- Swift, H. The constancy of desoxyribose nucleic acid in plant nuclei. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 643–654. [[CrossRef](#)]
- Fay, M.F.; Cowan, R.S.; Leitch, I.J. The effects of nuclear, DNA content (C-value) on the quality and utility of AFLP fingerprints. *Ann. Bot.* **2005**, *95*, 237–246. [[CrossRef](#)]
- Leitch, I.J.; Bennett, M.D. Genome size and its uses: The impact of flow cytometry. In *Flow Cytometry with Plant Cells: Analysis of Genes, Chromosomes and Genomes*; Wiley-VCH: Weinheim, Germany, 2007; pp. 153–176.
- Gregory, T.R. Genome size evolution in animals. In *The Evolution of the Genome*; Elsevier: Amsterdam, The Netherlands, 2005; pp. 3–87.
- Leong-Škorničková, J.; Šída, O.; Jarolímová, V.; Sabu, M.; Fér, T.; Trávníček, P.; Suda, J. Chromosome numbers and genome size variation in Indian species of *Curcuma* (Zingiberaceae). *Ann. Bot.* **2007**, *100*, 505–526. [[CrossRef](#)] [[PubMed](#)]
- Kelly, L.J.; Leitch, A.R.; Fay, M.F.; Renny-Byfield, S.; Pellicer, J.; Macas, J.; Leitch, I.J. Why size really matters when sequencing plant genomes. *Plant. Ecol. Divers.* **2012**, *5*, 415–425. [[CrossRef](#)]
- Doležel, J.; Greilhuber, J.; Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2007**, *2*, 2233. [[CrossRef](#)] [[PubMed](#)]
- Sliwinska, E. Flow cytometry—A modern method for exploring genome size and nuclear DNA synthesis in horticultural and medicinal plant species. *Folia Hort.* **2018**, *30*, 103–128. [[CrossRef](#)]
- Hanrahan, S.J.; Johnston, J.S. New genome size estimates of 134 species of arthropods. *Chromosome Res.* **2011**, *19*, 809–823. [[CrossRef](#)]
- Li, X.; Waterman, M.S. Estimating the repeat structure and length of DNA sequences using ℓ -tuples. *Genome Res.* **2003**, *13*, 1916–1922.
- Guo, L.T.; Wang, S.L.; Wu, Q.J.; Zhou, X.G.; Xie, W.; Zhang, Y.J. Flow cytometry and K-mer analysis estimates of the genome sizes of *Bemisia tabaci* B and Q (Hemiptera: Aleyrodidae). *Front. Physiol.* **2015**, *6*, 144. [[CrossRef](#)] [[PubMed](#)]
- He, K.; Lin, K.; Wang, G.; Li, F. Genome Sizes of Nine Insect Species Determined by Flow Cytometry and k-mer Analysis. *Front. Physiol.* **2016**, *7*, 569. [[CrossRef](#)] [[PubMed](#)]
- Li, R.; Fan, W.; Tian, G.; Zhu, H.; He, L.; Cai, J.; Huang, Q.; Cai, Q.; Li, B.; Bai, Y.; et al. The sequence and de novo assembly of the giant panda genome. *Nature* **2010**, *463*, 311–317. [[CrossRef](#)]
- Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **2011**, *475*, 189. [[CrossRef](#)]

18. Martin-Bravo, S.; Meimberg, H.; Luceno, M.; Markl, W.; Valcarcel, V.; Brauchler, C.; Vargas, P.; Heubl, G. Molecular systematics and biogeography of Resedaceae based on ITS and trnL-F sequences. *Mol. Phylogenet. Evol.* **2007**, *44*, 1105–1120. [[CrossRef](#)] [[PubMed](#)]
19. Yildirim, H.; Şenol, S.G. Reseda malatyana (Resedaceae), a new chasmophytic species from eastern Anatolia, Turkey. *Turk. J. Bot.* **2014**, *38*, 1013–1021. [[CrossRef](#)]
20. Sales, A.; Bagherizadeh, Y.; Malekzadeh, P.; Ahmadi, B.; Bonab, F. Evaluation of the antimicrobial effects of essential oil of reseda lutea L on pathogenic bacteria: Staphylococcus aureus, Staphylococcus epidermidis, and Escherichia coli. *Arch. Clin. Microbiol.* **2017**, *8*, 1–6. [[CrossRef](#)]
21. Casetti, F.; Jung, W.; Wolfle, U.; Reuter, J.; Neumann, K.; Gilb, B.; Wahling, A.; Wagner, S.; Merfort, I.; Schempp, C.M. Topical application of solubilized Reseda luteola extract reduces ultraviolet B-induced inflammation in vivo. *J. Photochem. Photobiol. B* **2009**, *96*, 260–265. [[CrossRef](#)] [[PubMed](#)]
22. Berrehal, D.; Khalfallah, A.; Bencharif-Betina, S.; Kabouche, Z.; Kacem, N.; Kabouche, A.; Calliste, C.-A.; Duroux, J.-L. Comparative antioxidant activity of two Algerian Reseda species. *Chem. Nat. Compd.* **2010**, *46*, 456–458. [[CrossRef](#)]
23. Ali, M.A.; Al-Hemaid, F.M.; Choudhary, R.K.; Lee, J.; Kim, S.-Y.; Rub, M. Status of Reseda pentagyna Abdallah & AG Miller (Resedaceae) inferred from combined nuclear ribosomal and chloroplast sequence data. *Bangladesh J. Plant. Taxon.* **2013**, *20*, 233–238.
24. Llewellyn, O.; Hall, M.; Miller, A.; Al-Abbasi, T.; Al-Wetaid, A.; Al-Harbi, R.; Al-Shammari, K.; Al-Farhan, A. Important Plant Areas in the Arabian Peninsula: 1. Jabal Qaraqir. *Edinb. J. Bot.* **2010**, *67*, 37. [[CrossRef](#)]
25. Abdallah, M.S. *The Resedaceae: A Taxonomical Revision of the Family*; De Landbouwhogeschool te Wageningen: Wageningen, Germany, 1967.
26. Alwadi, H.; Moustafa, F. Altitudinal impact on the weeds species distribution in the semi-arid mountainous region of Abha, Saudi Arabia. *J. Crop. Weed* **2016**, *12*, 87–95.
27. Pagnotta, E.; Montaut, S.; Matteo, R.; Rollin, P.; Nuzillard, J.-M.; Lazzeri, L.; Bagatta, M. Glucosinolates in Reseda lutea L: Distribution in plant tissues during flowering time. *Biochem. Syst. Ecol.* **2020**, *90*, 104043. [[CrossRef](#)]
28. Chaudhary, S. Resedaceae. In *Flora of the Kingdom of Saudi Arabia*; Chaudhary, S., Ed.; Ministry of Agriculture and Water, National Herbarium, National Agriculture and Water Research Center: Riyadh, Saudi Arabia, 1999; pp. 536–543.
29. Miller, A.; Nyberg, J. Studies in the Flora of Arabia: XXVII. Some new taxa from the Arabian Peninsula. *Edinb. J. Bot.* **1994**, *51*, 33–47. [[CrossRef](#)]
30. Murashige, T.; Skoog, F. A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant.* **1962**, *15*, 473–497. [[CrossRef](#)]
31. White, T.J.; Bruns, T.; Lee, S.; Taylor, J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protoc. A Guide Methods Appl.* **1990**, *18*, 315–322.
32. Gardes, M.; Bruns, T.D. ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **1993**, *2*, 113–118. [[CrossRef](#)] [[PubMed](#)]
33. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K.; Mega, X. Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [[CrossRef](#)]
34. Kumar, S.; Gadagkar, S.R. Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. *J. Mol. Evol.* **2000**, *51*, 544–553. [[CrossRef](#)]
35. Doležel, J.; Sgorbati, S.; Lucretti, S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Plant.* **1992**, *85*, 625–631. [[CrossRef](#)]
36. Sadhu, A.; Bhadra, S.; Bandyopadhyay, M. Novel nuclei isolation buffer for flow cytometric genome size estimation of Zingiberaceae: A comparison with common isolation buffers. *Ann. Bot.* **2016**, *118*, 1057–1070. [[CrossRef](#)] [[PubMed](#)]
37. Yan, J.; Zhang, J.; Sun, K.; Chang, D.; Bai, S.; Shen, Y.; Huang, L.; Zhang, J.; Zhang, Y.; Dong, Y. Ploidy level and DNA content of *Erianthus arundinaceus* as determined by flow cytometry and the association with biological characteristics. *PLoS ONE* **2016**, *11*, e0151948. [[CrossRef](#)]
38. Bennett, M.D.; Bhandol, P.; Leitch, I.J. Nuclear DNA amounts in angiosperms and their modern uses—807 new estimates. *Ann. Bot.* **2000**, *86*, 859–909. [[CrossRef](#)]
39. Andrews, S. FastQC V0.11.9: A quality control tool for high throughput sequence data. In *Babraham Bioinformatics*; Babraham Institute: Cambridge, UK, 2019. Available online: www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed on 10 April 2021).
40. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)] [[PubMed](#)]
41. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **2012**, *1*, 18. [[CrossRef](#)]
42. Xu, H.; Luo, X.; Qian, J.; Pang, X.; Song, J.; Qian, G.; Chen, J.; Chen, S. FastUniq: A fast de novo duplicates removal tool for paired short reads. *PLoS ONE* **2012**, *7*, e52249. [[CrossRef](#)]
43. Bushnell, B. BMAP: A fast, accurate, splice-aware aligner. In Proceedings of the 9th Annual, Genomics of Energy & Environment Meeting, Walnut Creek, CA, USA, 17–20 March 2014; Lawrence Berkeley National Lab: Berkeley, CA, USA, 2014.
44. Wingett, S.W.; Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **2018**, *7*, 1338. [[CrossRef](#)]

45. Chen, W.; Hasegawa, D.K.; Arumuganathan, K.; Simmons, A.M.; Wintermantel, W.M.; Fei, Z.; Ling, K.S. Estimation of the Whitefly *Bemisia tabaci* Genome Size Based on k-mer and Flow Cytometric Analyses. *Insects* **2015**, *6*, 704–715. [[CrossRef](#)]
46. Liu, B.; Shi, Y.; Yuan, J.; Hu, X.; Zhang, H.; Li, N.; Li, Z.; Chen, Y.; Mu, D.; Fan, W. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* **2013**, arXiv:1308.2012.
47. Marcais, G.; Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **2011**, *27*, 764–770. [[CrossRef](#)] [[PubMed](#)]
48. Chikhi, R.; Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **2014**, *30*, 31–37. [[CrossRef](#)]
49. Vurture, G.W.; Sedlazeck, F.J.; Nattestad, M.; Underwood, C.J.; Fang, H.; Gurtowski, J.; Schatz, M.C. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **2017**, *33*, 2202–2204. [[CrossRef](#)]
50. Sun, H.; Ding, J.; Piednoël, M.; Schneeberger, K. findGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **2018**, *34*, 550–557. [[CrossRef](#)] [[PubMed](#)]
51. Hozza, M.; Vinař, T.; Brejová, B. How big is that genome? Estimating genome size and coverage from k-mer abundance spectra. In *International Symposium on String Processing and Information Retrieval: 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 199–209.
52. Lander, E.S.; Waterman, M.S. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **1988**, *2*, 231–239. [[CrossRef](#)]
53. Ryan, J. Estimate_Genome_Size. PI (Version 0.03) [Computer Software]. Sars International Centre for Marine Molecular Biology, Bergen, Norway 2013. Available online: http://josephryan.github.com/estimate_genome_size.pl/ (accessed on 10 April 2021).
54. Ranallo-Benavidez, T.R.; Jaron, K.S.; Schatz, M.C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **2020**, *11*, 1432. [[CrossRef](#)] [[PubMed](#)]
55. Loureiro, J.; Doležel, J.; Greilhuber, J.; Santos, C.; Suda, J. Plant Flow Cytometry—Far beyond the Stone Age. *Cytom. Part A* **2008**, *73a*, 579–580. [[CrossRef](#)] [[PubMed](#)]
56. Knight, C.A.; Ackerly, D.D. Variation in nuclear DNA content across environmental gradients: A quantile regression analysis. *Ecol. Lett.* **2002**, *5*, 66–76. [[CrossRef](#)]
57. Beaulieu, J.M.; Moles, A.T.; Leitch, I.J.; Bennett, M.D.; Dickie, J.B.; Knight, C.A. Correlated evolution of genome size and seed mass. *New Phytol.* **2007**, *173*, 422–437. [[CrossRef](#)]
58. Knight, C.A.; Beaulieu, J.M. Genome size scaling through phenotype space. *Ann. Bot.* **2008**, *101*, 759–766. [[CrossRef](#)]
59. Knight, C.A.; Molinari, N.A.; Petrov, D.A. The large genome constraint hypothesis: Evolution, ecology and phenotype. *Ann. Bot.* **2005**, *95*, 177–190. [[CrossRef](#)] [[PubMed](#)]
60. Vinogradov, A.E. Selfish DNA is maladaptive: Evidence from the plant Red List. *Trends Genet.* **2003**, *19*, 609–614. [[CrossRef](#)]
61. Vidic, T.; Greilhuber, J.; Vilhar, B.; Dermastia, M. Selective significance of genome size in a plant community with heavy metal pollution. *Ecol. Appl.* **2009**, *19*, 1515–1521. [[CrossRef](#)]
62. Doležel, J.; Greilhuber, J.; Lucretti, S.; Meister, A.; Lysák, M.; Nardi, L.; Obermayer, R. Plant genome size estimation by flow cytometry: Inter-laboratory comparison. *Ann. Bot.* **1998**, *82* (Suppl. 1), 17–26. [[CrossRef](#)]
63. Soltis, D.E.; Soltis, P.S.; Bennett, M.D.; Leitch, I.J. Evolution of genome size in the angiosperms. *Am. J. Bot.* **2003**, *90*, 1596–1603. [[CrossRef](#)] [[PubMed](#)]
64. González-Aguilera, J.; Fernández-Peralta, A.M. Phylogenetic relationships in the family Resedaceae L. *Genetica* **1984**, *64*, 185–197. [[CrossRef](#)]
65. Eigsti, O.J. Cytological studies in the Resedaceae. *Bot. Gaz.* **1936**, *98*, 363–369. [[CrossRef](#)]
66. Michael, T.P.; VanBuren, R. Progress, challenges and the future of crop genomes. *Curr. Opin. Plant. Biol.* **2015**, *24*, 71–81. [[CrossRef](#)]
67. Williams, D.; Trimble, W.L.; Shilts, M.; Meyer, F.; Ochman, H. Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. *BMC Genom.* **2013**, *14*, 1–11. [[CrossRef](#)]
68. Gregory, T.R.; Nathwani, P.; Bonnett, T.R.; Huber, D.P. Sizing up arthropod genomes: An evaluation of the impact of environmental variation on genome size estimates by flow cytometry and the use of qPCR as a method of estimation. *Genome* **2013**, *56*, 505–510. [[CrossRef](#)]
69. DeSalle, R.; Gregory, T.R.; Johnston, J.S. Preparation of samples for comparative studies of arthropod chromosomes: Visualization, in situ hybridization, and genome size estimation. *Methods Enzym.* **2005**, *395*, 460–488.
70. Hardie, D.C.; Gregory, T.R.; Hebert, P.D. From pixels to picograms: A beginners' guide to genome quantification by Feulgen image analysis densitometry. *J. Histochem. Cytochem.* **2002**, *50*, 735–749. [[CrossRef](#)] [[PubMed](#)]
71. Zhang, G.; Fang, X.; Guo, X.; Li, L.; Luo, R.; Xu, F.; Yang, P.; Zhang, L.; Wang, X.; Qi, H.; et al. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **2012**, *490*, 49–54. [[CrossRef](#)] [[PubMed](#)]
72. Edwards, R.J.; Tuipulotu, D.E.; Amos, T.G.; O'Meally, D.; Richardson, M.F.; Russell, T.L.; Vallinoto, M.; Carneiro, M.; Ferrand, N.; Wilkins, M.R.; et al. Draft genome assembly of the invasive cane toad, *Rhinella marina*. *Gigascience* **2018**, *7*, giy095. [[CrossRef](#)]
73. Hu, Y.; Resende, M.F.; Bombarely, A.; Brym, M.; Bassil, E.; Chambers, A.H. Genomics-based diversity analysis of *Vanilla* species using a *Vanilla planifolia* draft genome and Genotyping-By-Sequencing. *Sci. Rep.* **2019**, *9*, 1–16. [[CrossRef](#)] [[PubMed](#)]
74. Hedgecock, D.; Gaffney, P.M.; Gouletquer, P.; Guo, X.; Reece, K.; Warr, G.W. The case for sequencing the Pacific oyster genome. *J. Shellfish. Res.* **2005**, *24*, 429–441.

75. Elliott, T.A.; Gregory, T.R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. B Biol. Sci.* **2015**, *370*, 20140331. [[CrossRef](#)]
76. Jansen, H.J.; Liem, M.; Jong-Raadsen, S.A.; Dufour, S.; Weltzien, F.A.; Swinkels, W.; Koelewijn, A.; Palstra, A.P.; Pelster, B.; Spaink, H.P.; et al. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Sci. Rep.* **2017**, *7*, 7213. [[CrossRef](#)]
77. Jimenez, A.G.; Kinsey, S.T.; Dillaman, R.M.; Kapraun, D.F. Nuclear DNA content variation associated with muscle fiber hypertrophic growth in decapod crustaceans. *Genome* **2010**, *53*, 161–171. [[CrossRef](#)]
78. Kidwell, M.G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **2002**, *115*, 49–63. [[CrossRef](#)]
79. Li, S.F.; Su, T.; Cheng, G.Q.; Wang, B.X.; Li, X.; Deng, C.L.; Gao, W.J. Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes* **2017**, *8*, 290. [[CrossRef](#)]
80. Baeza, J.A.; MacManes, M. De novo assembly and functional annotation of the heart + hemolymph transcriptome in the Caribbean spiny lobster *Panulirus argus*. *Mar. Genom.* **2020**, *54*, 100783. [[CrossRef](#)]
81. SanMiguel, P.; Tikhonov, A.; Jin, Y.K.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P.S.; Edwards, K.J.; Lee, M.; Avramova, Z.; et al. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **1996**, *274*, 765–768. [[CrossRef](#)]
82. Chandler, V.L.; Brendel, V. The Maize Genome Sequencing Project. *Plant. Physiol.* **2002**, *130*, 1594–1597. [[CrossRef](#)] [[PubMed](#)]
83. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338–345. [[CrossRef](#)] [[PubMed](#)]
84. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genom. Proteom. Bioinform.* **2015**, *13*, 278–289. [[CrossRef](#)] [[PubMed](#)]
85. Rang, F.J.; Kloosterman, W.P.; de Ridder, J. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **2018**, *19*, 90. [[CrossRef](#)] [[PubMed](#)]
86. Pellicer, J.; Leitch, I.J. The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. *New Phytol.* **2020**, *226*, 301–305. [[CrossRef](#)]
87. Rice, A.; Glick, L.; Abadi, S.; Einhorn, M.; Kopelman, N.M.; Salman-Minkov, A.; Mayzel, J.; Chay, O.; Mayrose, I. The Chromosome Counts Database (CCDB)—A community resource of plant chromosome numbers. *New Phytol.* **2015**, *206*, 19–26. [[CrossRef](#)] [[PubMed](#)]
88. Goldblatt, P.; Johnson, D. *Index to Plant Chromosome Numbers (ICPN Reports)*; Missouri Botanical Garden: St. Louis, Missouri, 2015.
89. Gonzhlez-Aguilera, J.; FernCtndez-Peralta, A.; Safiudo, A. Cytogenetic and evolutionary studies on the Spanish species of the Family Resedaceae L: Sections *Phyteuma* L and *Resedastrum* Duby. *Bol. Soc. Brot. Ser.* **1980**, *2*, 519–536.
90. Pustahija, F.; Brown, S.C.; Bogunić, F.; Bašić, N.; Muratović, E.; Ollier, S.; Hidalgo, O.; Bourge, M.; Stevanović, V.; Siljak-Yakovlev, S. Small genomes dominate in plants growing on serpentine soils in West Balkans, an exhaustive study of 8 habitats covering 308 taxa. *Plant. Soil* **2013**, *373*, 427–453. [[CrossRef](#)]
91. Sañudo, A.; Rejon, R. Sobre la Naturaleza Autoploide de Algunas Plantas Silvestre. *An. Inst. Bot. Cavanilles* **1975**, *32*, 633–648.