 doi:10.20944/preprints202204.0255.v1

Article

First Genomic Insights Into The *Mandevilla* Genus

Fabio Palumbo [1], Samela Draga [1], Francesco Scariolo [1], Gio Batta Sacilotto [2], Marco Gazzola [2] and Gianni Barcaccia [1,*]

[1] Department of Agronomy Food Natural resources Animals Environment, Campus of Agripolis, University of Padova, 35020 Legnaro, Padova Italy.
[2] Gruppo Padana Ortofloricoltura S.S., Via Olimpia 41, 31038, Paese, Treviso, Italy
* Correspondence: gianni.barcaccia@unipd.it

Abstract: *Mandevilla* (Apocynaceae) is a greatly appreciated genus in the world ornamental market. In this study, we attempted to address the poor genetic knowledge and the huge taxonomic gaps existing in this genus by analyzing the germplasm of 55 accessions. After cytometrically determining the triploid genome size (1,512.64 Mb) of a reference sample (variety "Mandevilla 2001"), the plastidial genome (cpDNA, 0,18 Mb) and a draft of the nuclear genome (nuDNA, 207 Mb) were assembled. While cpDNA was effective in reconstructing the phylogenesis of the Apocynaceae family based on a DNA superbarcoding approach, the nuDNA assembly length was found to be only 41% of the haploid genome size (506 Mb, predicted based on the K-mer frequency distribution). Its annotation enabled the prediction of thousands of genes, of which 5,275 resulted in full CDS. Among them, we identified nine genes whose orthologs (in *Catharanthus roseus*) encode enzymes involved in the biosynthesis of monoterpene indole alkaloids (MIAs), including catharanthine, tabersonine and vincadifformine. The nuclear genome draft was also useful to develop a highly informative (mean PIC=0.62) set of 23 SSR markers that was validated on the Mandevilla collection. These results were integrated with cytometric measurements, nuclear ITS1 haplotyping and chloroplast DNA barcoding analyses to assess the origin, divergence and relationships existing among the 55 accessions object of the study. As expected, based on the scarce information available in the literature, the scenario was extremely intricate. A reasonable hypothesis is that most of the accessions represent interspecific hybrids sharing the same species as maternal parent (i.e., *Mandevilla sanderi*).

Keywords: genome assembly; Apocynaceae; SSR; superbarcoding; flow cytometry; monoterpene indole alkaloids; DNA barcoding; cpDNA; *Mandevilla sanderi*

1. Introduction

Mandevilla Lindl. (in honor of the diplomat Henry Mandeville) is a genus of plants greatly appreciated for their pink, white or red flowers, whose blooming time usually fluctuates from late spring to autumn. It is native to Middle and South America [1] and belongs to the Apocynaceae family, along with more than 4600 species [2]. Several Apocynaceae family members are highly reputed in traditional medicine and have been extensively investigated for a wide range of curative properties, including antioxidant [3], anti-inflammatory [2], anticancer [4] and antimicrobial [5] activities. Mandevilla instead represents a leading product of the world ornamental scenario, as demonstrated by the growing number of new varieties that literally flood the market every year [6]. Despite this, very few studies have been accomplished in this species, and many biological questions remain unsolved. Starting from the genus name, the debate about the use of *Dipladenia* (meaning "with two glands") as a synonym of *Mandevilla* is still

 doi:10.20944/preprints202204.0255.v1

open. A few years after the first classification of the *Mandevilla* genus (1840), Alphonse de Candolle published in Vol. VIII of *Prodromus Systematis Naturalis Regni Vegetabilis* a revision of Apocynaceae systematics, establishing a new genus called *Dipladenia*, different from the sister genus *Mandevilla* [7]. Almost 100 years later, a second classification [8] revolutionized the genus organization: *Dipladenia* spp., along with seven other genera, all of which were incorporated in the *Mandevilla* genus. Although Dipladenia is currently generally considered a historical synonym of Mandevilla, a recent study demonstrated that the Dipladenia group clustered apart from the Mandevilla group [6]. Moreover, some breeding companies are still tied to tradition, and they continue selling their products, clearly distinguishing Mandevilla from Dipladenia [9]. In fact, although from a botanical point of view this distinction is considered inconsistent, from a commercial point of view, the term Mandevilla is sometimes still used to indicate vigorous climbing varieties with large leaves, while Dipladenia is sometimes used for more compact and bushy varieties with small leaves.

In addition to these intergenus disputes, intragenus classification has also faced significant complications. The number of species included in the *Mandevilla* genus increased from 108 (from a first systematic classification, in 1933 [8]) to 179 (as reported by the Royal Botanic Gardens, Kew [10]), but thus far, only 50 have been molecularly characterized and registered in BOLD systems and NCBI. Except for Simões et al. [11], who analyzed 47 different *Mandevilla* species, exhaustive and elucidating DNA barcoding-based studies are missing in this genus. What makes this genus challenging for taxonomists is the remarkable morphological variation observed among and within species due to a continuous process of adaptation to a plethora of different geographical environments [11]. Another level of complexity is given by the fact that most of the genotypes available are interspecific hybrids, characterized by complex genealogies [6] and possibly even by different levels of ploidy or by unbalanced numbers of chromosomes. Additionally, Mandevilla is an understudied species: no information about DNA content (c-value) and ploidy levels is available in the scientific literature.

Taxonomic classification and phylogenetic relationships are not the only lacking aspects; molecular markers, with one exception [6], have never been exploited in these species, suggesting that despite the huge turnover, breeding schemes are still planned almost exclusively on a phenotypic basis. Historically, marker-assisted breeding in ornamental species has been delayed compared to crop species for several reasons, such as reduced economic importance, genome complexity (i.e., ornamental plants have genomes that are usually larger and polyploid) and methods of propagation (i.e., ornamental plants are mainly vegetatively propagated). More recently, the collapse of sequencing prices has made next-generation sequencing platforms accessible to noncrop species, including ornamental plants. This contributed to the development of codominant molecular markers (i.e., SSR and SNP) in species such rose [12], lilium [13], tulip [14] or chrysanthemum [15].

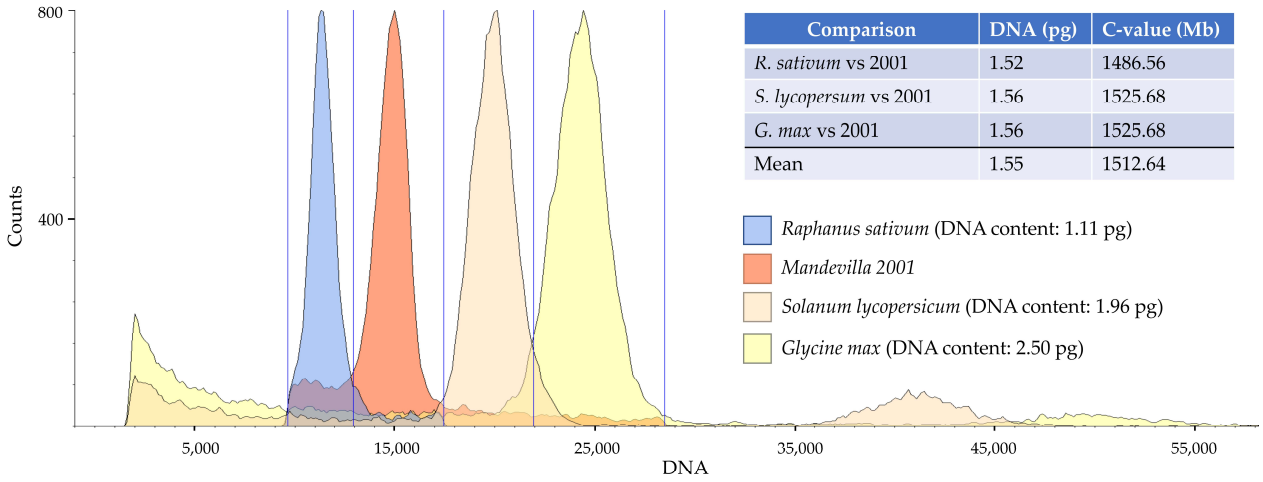
To take a first step in this direction, we tried to address some of the main deficits existing in this genus from both a cytometric and genomic point of view. In the first case, we first estimated the size of the genome and the DNA content of some samples commercially available on the market. In the second case, we developed and annotated a first draft of the genome with the identification of thousands of genes and markers useful for marker-assisted breeding. Notably, we assembled plastid DNA and assessed its effectiveness in taxonomy classification through DNA superbarcoding analysis.

2. Results and discussion

 doi:10.20944/preprints202204.0255.v1

2.1 Genome size estimate through flow citometry

Genome size data for the Apocynaceae family are quite scant: only 35 species (of approximately 4,600) have been cytometrically investigated, and the *Mandevilla* genus is not among them [16]. In this study, the genome size of "Mandevilla 2001" was successfully determined by costaining its nuclei along with those extracted from *Raphanus sativus*, *Glycine max* and *Solanum lycopersicum*. These reference standards were chosen because the sizes of their genomes are placed in a range (from 1,085 Mb for *R. sativus* to 2,446 Mp for *G. max*) sufficiently wide to cover the putative genome size of *Mandevilla* spp. In fact, although data about the DNA content of this genus are lacking, the genome size for most of the known Apocynaceae species ranges from 490 Mb (*Apocynum androsaemifolium*) to 2,322 Mb (*Vinca difformis*). Notably, despite the usage of three references from as many families, estimates of 2001 sample genome size provided highly overlapping values, ranging from 1.52 pg to 1.56 pg (**Figure 1**). The slight difference (3.4%) observed among the measurements, according to Bennett et al., [17] could be due to the presence of inhibitors (e.g., anthocyanin) preventing PI binding to DNA or variability in chromatin condensation across samples. Overall, assuming 1 pg = 0.978 Gbp [18], the genome size of the Mandevilla sample oscillates between 1,486.56 and 1,525.68 Mb.



Comparison	DNA (pg)	C-value (Mb)
R. sativum vs 2001	1.52	1486.56
S. lycopersum vs 2001	1.56	1525.68
G. max vs 2001	1.56	1525.68
Mean	1.55	1512.64

Figure 1. Flow cytometry analysis for genome size estimates. Each peak represents the total DNA fluorescence emission of propidium iodide (PI)-stained leaf nuclei purified (in order from left to right) from *Raphanus sativum*, Mandevilla 2001, *Solanum lycopersicum*, *Gycine max*.

2.2 Sequencing output, organelle and genome assembly

Whole-genome sequencing yielded 314,414,127 x 2 reads (~62.8 Gb); 97.3% of them had a Phred quality score of 20, and 91.7% had a Phred quality score of Q30. After adapter removal, a quality check (Phred Quality Score >30) and a trimming step (read length > 40 bp), 56.41 Gb were retained and used for both plastid and nuclear genome assembly. The chloroplast genome assembly (deposited in GenBank under accession ID OM489306), with a length of 180,004, resulted in the largest plastid organelle assembled from its family. In fact, the 33 cpDNA available in NCBI for the Apocynaceae family have a size ranging from 153,826 (*Periploca forrestii*) to 176,340 (*Hoya carnosa*). The chloroplast genome showed a small single copy (SSC), a large single copy (LSC) and two inverted repeats (IRA and IRB, **Figure 2**, panel A).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0255.v1

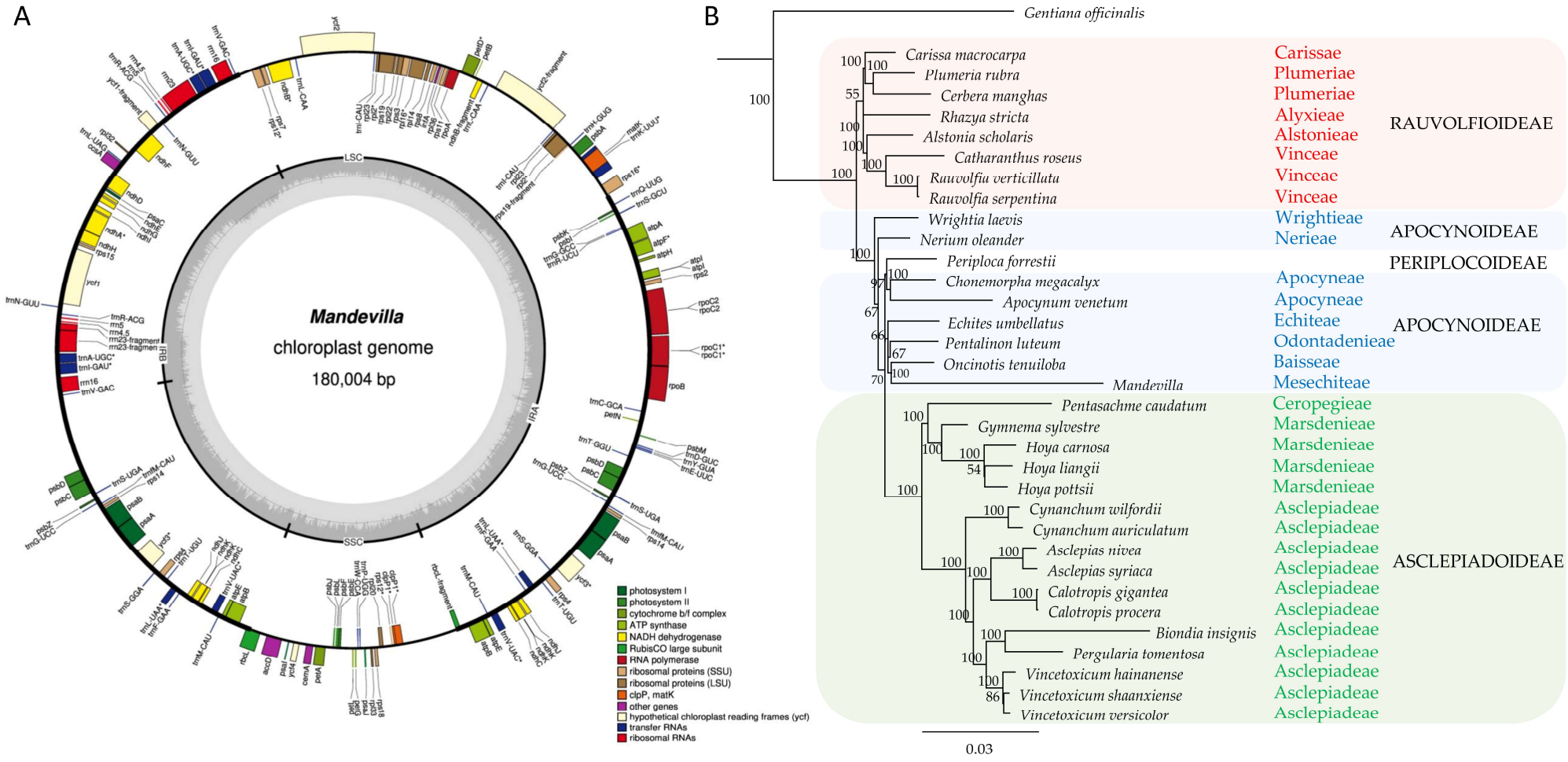


Figure 2. (**A**) Annotation of the Mandevilla 2001 plastidial genome. Genes marked with an asterisk (*) contain introns. (**B**) The newly assembled Mandevilla plastome was aligned using MAFFT v7.450 under default parameters with 33 plastid genomes (from the Apocynaceae family) retrieved from GenBank; Geneious prime (Biomatters, Inc., San Diego, CA, USA) was used to reconstruct the phylogenetic relationships based on the neighbor-joining (NJ) method, and 1,000 bootstrap replicates. *Gentiana officinalis* (Gentianaceae family) was used as an outgroup.

 doi:10.20944/preprints202204.0255.v1

These latter have been reported, with few exceptions, in almost all land plants and are considered an ancestral feature that was lost independently in several plant families [19]. Overall, the genome contained 107 different genes, including 73 protein-coding, 30 tRNA, and 4 rRNAgenes (**Supplementary Table 1**). Among them, 10 genes contained one intron, and 6 genes (*trn*V-UAC, *trn*I-GAU, *trn*A-UGC, *trn*L-UAA, *rpl2*, and *ycf3*) contained two introns. The newly assembled plastome was used to investigate the phylogenetic relationship between the *Mandevilla* genus and 34 other Apocynaceae species based on a DNA superbarcoding approach. From the plastome-based tree, all species were correctly grouped into their 15 relative tribes with bootstrap values usually equal to 100 (**Figure 2**, panel B). Mandevilla, which is part of the Mesechiteae tribe, clustered with all tribes belonging to the Apocynoideae subfamily. Moreover, the relationships existing among the four subfamilies (Rauvolfioideae, Apocynoideae, Periplocoidae and Asclepiadoideae) perfectly matched those reported by Livshultz [20] and Livshultz et al., [21] although the analyses were conducted differently (i.e., aligning entire plastomes and single plastidial genes such as *trn*L intron, *trn*L-*trn*F spacer, *rpl16* intron, and *mat*K).

Regarding the nuclear DNA assembly, filtered and merged paired-end data were assembled into 116,244 contigs encompassing a total of 207,657,661 Mb (**Table 1**). The genome draft was deposited in GenBank under accession ID PRJNA802340.

Table 1. Assembly statistics and BUSCO analysis results

Assembly statistics	
N. contigs	116,244
Total length (bp)	207,657,661
N. contigs > 500 bp	86,428
Largest contig (bp)	36,523
Total length (bp; > 500 bp)	197,539,476
GC (%)	38.38
N50 > 500 bp	3335
L50 > 500 bp	16714
Coverage	62 ×
BUSCO analysis (2326 total BUSCO groups)	
Complete	835 (816 single copy)
Fragmented	470

Additionally, based on the K-mer frequency distribution analysis (K=25), the haploid genome size was estimated to be 506 Mb. According to the genome size estimates obtained using flow cytometry (~1,512 Mb), it was possible to predict the triploidy of the Mandevilla sample. Moreover, comparing the assembled genome and the estimated haploid genome, only 41% of the genome was assembled. This is in agreement with the BUSCO analysis findings: the assembly retrieved 55.3% of the conserved single-copy ortholog genes, including 35.1% complete and 20.2% partial genes (**Table 1**). Out of 116,244 contigs, 86,428 had a size ranging from 500 to 36,523 bp (N50 = 3335 bp; L50=37,125, **Table 1**). Finally, by applying the Lander/Waterman equation, the average coverage was found to be approximately 62×.

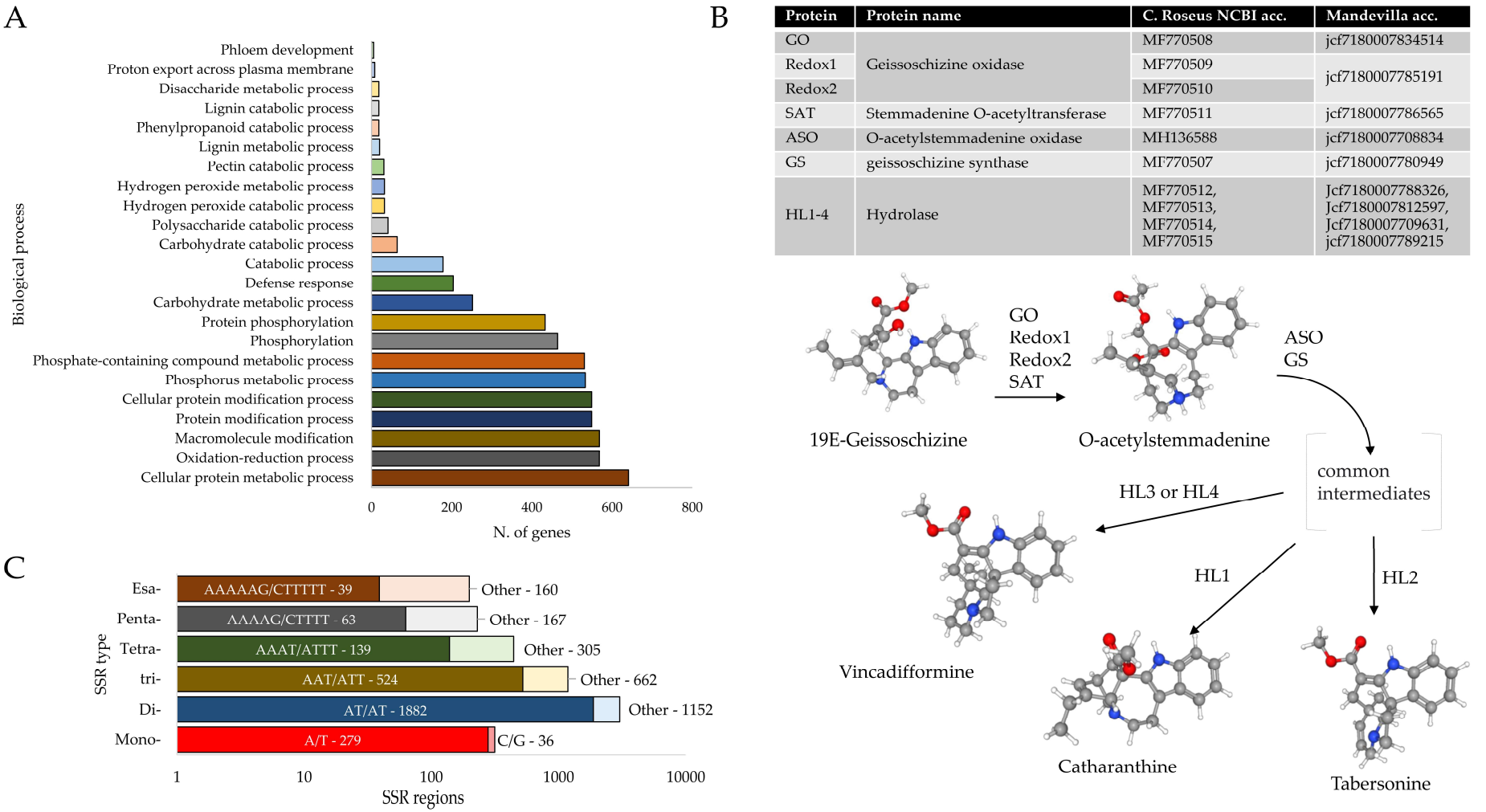
2.3 Genome annotation, SSR screening and validation

From the tBLASTn analysis (E-value ≤1e-10, BLAST+ v.2.3.0), 22,251 proteins (out of 25,574) from the *C. canephora* proteome showed a significant match (E-value ≤1e-10, **Supplementary Table 2**) in 13,371 contigs. In particular, the full

 doi:10.20944/preprints202204.0255.v1

CDSs of 5,275 proteins (20.62% of the entire *C. canephora* proteome) were orthologous within the newly assembled genome. Among them, 2,689 (50.98%) exhibited similarity scores higher than 60% and 2,005 (38.01%) had E-values <10e-100. The gene list codifying for the 5,275 proteins was linked with underlying molecular pathways by using gene ontology (GO) categories. The most enriched biological process categories in terms of the absolute number of genes were cellular protein metabolic process (641 genes), oxidation–reduction process (568 genes) and macromolecule modification (568 genes), while the most enriched categories in terms of completeness were phloem development (5 out of 5), proton export across the plasma membrane (8 out of 9) and disaccharide metabolic process (18 of 37; **Figure 3**, panel A). Among the full-length CDSs identified within the newly assembled genome, we searched for sequences putatively codifying for enzymes involved in the production of monoterpene indole alkaloids (MIAs). MIAs represent a large group of compounds mainly isolated from three plant families (the Loganiaceae, Apocynaceae, and Rubiaceae) and classified based on the geometric arrangement of the C9/C10 carbon skeleton in the three main configurations: Corynanthe, Aspidosperma, and Iboga types [22]. Isolation in *Catharanthus roseus* (Madagascar periwinkle, Apocynaceae) of two MIAs (vinblastine and vincristine) able to block cell division is considered a milestone in cancer chemotherapy [23]. Since then, several MIA-based drugs have been under consideration for different types of cancer treatment. Due to the great interest behind the production of these molecules, several studies have recently attempted to elucidate the metabolic pathway leading to the production of catharanthine (iboga MIA), tabersonine (aspidosperma MIA) and vincadifformine (aspidosperma MIA) [23–25]. The first two represent precursors of vinblastine, while the synthesis of the last compound appears to branch off prior to the formation of tabersonine (**Figure 3**, panel B). Although the identification of these MIAs has been demonstrated in *C. roseus*, it has yet to be elucidated how many other Apocynaceae species are capable of producing this class of molecules. In a recent study, 444 IMAs, including catharanthine, vincadifformine and tabersonine, were identified in six genera belonging to this family (*Alstonia*, *Rauvolfia*, *Kopsia*, *Ervatamia*, *Tabernaemontana*, and *Rhazya*) [26]. However, the production of MIAs in the *Mandevilla* genus has never been proven or investigated. From the tBLASTn analysis performed by aligning 10 amino acid sequences involved in the biosynthesis of catharanthine, tabersonine and vincadifformine in *C. roseus* (Apocynaceae) against the Mandevilla genome, we identified (E-values ranging from 2.28E-69 and 1.27E-144) the full CDS of two putative geissoschizine oxidases, a stemmadenine O-acetyltransferase, an O-acetylstemmadenine oxidase, a geissoschizine synthase and four putative hydrolases (**Figure 3**, panel B, and **Supplementary Table 3**). The latter, orthologous to HL1-HL4, could be specifically involved in the last step of biosynthesis of the three abovementioned MIA compounds. However, considering the high similarity values shared by these four HL genes, we were not able to discriminate whether a specific sequence that encoded a hydrolase was involved in an IMA pathway (e.g., tabersonine) rather than another (e.g., catharanthine). qPCR analysis on target genes conducted in different tissues and at different stages of development [27] conjugated with liquid chromatography–mass spectrometry-based analyses will be crucial to elucidate if and possibly which MIA *Mandevilla* genus is able to produce.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0255.v1



Protein	Protein name	C. Roseus NCBI acc.	Mandevilla acc.
GO	Geissoschizine oxidase	MF770508	jcf7180007834514
Redox1		MF770509	jcf7180007785191
Redox2		MF770510	
SAT	Stemmadenine O-acetyltransferase	MF770511	jcf7180007786565
ASO	O-acetylstemmadenine oxidase	MH136588	jcf7180007708834
GS	geissoschizine synthase	MF770507	jcf7180007780949
HL1-4	Hydrolase	MF770512, MF770513, MF770514, MF770515	Jcf7180007788326, Jcf7180007812597, Jcf7180007709631, jcf7180007789215

Figure 3 (A) Most enriched biological process (BP) categories in terms of absolute number of genes. Enrichment analyses were conducted only taking into account those sequences (5,275) whose CDSs were complete (based on a tblastN alignment against the *C. canephora* proteome). **(B)** Identification of putative enzymes involved in the biosynthesis of catharanthine, tabersonine and vincadifformine. A tBLASTn analysis was performed by aligning 10 amino acid sequences from *C. roseus* involved in the production of the abovementioned MIAs (monoterpene indole alkaloids) against the Mandevilla genome. In the bottom part of the panel, the last steps leading to the production of catharanthine, tabersonine and vincadifformine are illustrated according to Qu et al. [23] **(C)** Detailed abundance of microsatellites in the newly assembled genome of Mandevilla. The most abundant motif for each SSR type is reported.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0255.v1

Predetermined MISA scripts identified 5.408 SSRs distributed over 4,806 contigs. The total length of the SSR-containing motifs detected in the genome draft was 66,135 Mb, barely 0.032% of the assembled sequences. This value, which is ten to thirty times lower than that observed in other plant species [28], could depend on both the stringent parameters used to search for SSRs (e.g., minimum repeat numbers of 20, 10 and 7 for mono-, di-, and trinucleotide repeat motifs, respectively) and the type of sequencing platform used. In fact, since genome assembly relies solely on Illumina sequencing, the exclusive use of short sequences can lead to an underestimation of repeated regions that, in many cases, cannot be properly assembled (unless the repeated region is shorter than the read length) [29]. The most frequent repeats were dinucleotides (56.10%) and trinucleotides (21.93%), while the most abundant dinucleotide and trinucleotide repeat motif types were AT/AT (62.03% of all dinucleotides) and AAT/ATT (44.18% of all trinucleotides; **Figure 3**, panel C). The fact that these motifs were the most abundant is fully in agreement with what was highlighted by Srivastava et al. in a recent SSR meta-analysis performed on 71 plant species [30].

From an initial panel of 100 SSRs (**Supplementary Table 4**), 23 SSRs (**Table 2**) were successfully employed to analyze, in multiplex, the 55-sample germplasm collection.

Table 2. Information on the 23 microsatellite markers validated using a collection of 55 Mandevilla samples

SSR	Marker Size	Primer F	Primer R	Mean T_m (°C)	Anchor	Multiplex
SSR_02	327	ATTTGTTTGCAACCTCCATG	CCGCAACTCAAACTCAAATT	55	PAN1	1
SSR_12	153	TGAAATAAAGGGTTAGGGCA	TCACTAATCCAGACAATCACA	54.2	PAN3	1
SSR_30	233	CAACACCTATACCTCACACC	GAGTTTGTAGTCTCCAACCTT	55.2	M13	1
SSR_34	202	TCTCCAATTAGCAGTACAAGG	TTAGACAGGGAGAGAGACAG	55	PAN1	1
SSR_41	171	GCCTCTCAAGTCATTAGGTG	AGGGTACTAAGGATGGTCTAA	55.5	PAN2	1
SSR_47	157	TGCTGCATTAATCACCTACA	GGCAGAAGAAGATTTGTCCA	54.6	M13	1
SSR_28	415	GAGATCAATGAGGATGGGAC	CACTTACAGTTTCAGGTCCT	54.6	M13	2
SSR_40	300	TGGACGAACTTGATACTACG	TGTTGAAAATCCCAGTCCAA	54.7	PAN1	2
SSR_50	138	CATTCAGCACACAGTTCTTC	AGTCATCGTTGTGAAATGGA	54.9	M13	2
SSR_15	223	TGAGGCACATACCATAGAGA	AATTTCTTGTCGTGGGCTAT	55	PAN3	2
SSR_48	187	CCGTGCCTCCTATGATTTAC	CTGACCATGCAATTACTCCT	55.1	M13	2
SSR_60	336	CCCTAGAGACCTTTTCATCC	CGAGTGTCTTCAAGCCATTA	54.5	M13	3
SSR_59	163	ATTCAGCACACAGTTCTTCT	GTCTATGACGGAGAGAAACC	54.9	M13	3
SSR_67	115	TACTAATTCGTCGTTTGGCT	CTTTTAGGTCATTTGGTCCAA	54	PAN1	3
SSR_55	185	TTTCAGCATAGGTTCGACAA	AAAGCCTGAATCTCCTCTTG	55	PAN2	3
SSR_76	272	AATAAACAGCCCAGTCTCAA	TTCTTCAATTTGCAGCCTTT	54.2	PAN3	3
SSR_89	424	AAACTGGGACCATACACATC	TTGACGTAACTGTTTGACCA	55	PAN3	3
SSR_74	253	GACGGATGCTCTTAATTCCT	GTGTACAGATCCCTACTTCC	54.3	M13	4
SSR_64	206	GGCACCTGTTAATATCAGTG	GATGGATGTAGAGGATGGTG	53.8	M13	4
SSR_70	345	TATTGAGGTTTGGCTTTCGA	CATTAACACCCCTCTTGTCA	55	PAN1	4
SSR_80	421	CTTTGGATTTGAAAGCGGAA	CAAAGGTATGTCTCTGGGTC	54.8	PAN2	4
SSR_61	413	ACAAAGCTTCTCCATCTCAG	GGGTGACTTTCCTGCTAATT	55.1	PAN3	4
SSR_95	139	ATTTTCCGTGAATCCAGATCT	TGAGAAGGGGTTGTTGTTG	55	PAN3	4

Descriptive statistics demonstrated a high polymorphic information content (PIC) for almost all the loci, with values ranging from 0.36 (SSR_28) to 0.82 (SSR_59) and an average value of 0.62. Based on the Botstein et al. classification [31], the majority of SSRs (18 loci, 78%) were highly polymorphic, with PIC values always higher than 0.5. Moreover, it is worth pointing out that the overall PIC of this new set of SSR panels was demonstrated to be higher than that of the only other SSR set available for the *Mandevilla* genus [6]. In this latter study, Oder

 doi:10.20944/preprints202204.0255.v1

et al. tested a set of 20 SSRs on 11 accessions belonging to the *Mandevilla* genus, and their PIC values oscillated from 0.08 to 0.76, and the average was 0.48. Looking at the number of observed (n_a) and effective (n_e) alleles, values ranged from 3 to 10 (mean 5.48) and from 1.55 to 4.99 (mean 2.95), respectively. Other parameters are presented in **Table 3.**

Table 3. SSR descriptive statistics reporting marker locus name, sample size of individuals successfully amplified for each locus, number of observed alleles (n_a), number of effective alleles (n_e), observed heterozygosity (H_o), expected heterozygosity (H_e), Shannon's information index (I) and polymorphic information content (PIC)

Marker	Sample size	n_a	n_e	H_o	H_e	I	PIC
SSR_02	53	3	2.04	0.51	0.51	0.74	0.51
SSR_12	52	4	2.98	0.88	0.67	1.18	0.66
SSR_15	45	5	3.72	0.91	0.74	1.41	0.73
SSR_28	47	3	1.55	0.30	0.36	0.60	0.36
SSR_30	45	4	1.78	0.20	0.44	0.78	0.44
SSR_34	50	7	3.88	0.76	0.75	1.59	0.74
SSR_40	46	6	3.66	0.67	0.73	1.41	0.73
SSR_41	50	10	4.18	0.70	0.77	1.69	0.76
SSR_47	55	6	1.64	0.36	0.39	0.86	0.39
SSR_48	49	5	2.11	0.43	0.53	0.95	0.53
SSR_50	51	9	4.99	0.90	0.81	1.80	0.80
SSR_55	55	4	3.32	0.84	0.71	1.27	0.70
SSR_59	55	8	4.92	0.93	0.80	1.75	0.82
SSR_60	40	3	1.81	0.08	0.45	0.68	0.45
SSR_61	53	6	2.39	0.43	0.59	1.20	0.58
SSR_64	55	4	2.15	0.60	0.54	0.87	0.53
SSR_67	54	4	1.83	0.37	0.46	0.83	0.45
SSR_70	53	6	3.59	0.66	0.73	1.50	0.72
SSR_74	52	5	3.15	0.46	0.69	1.30	0.68
SSR_76	54	6	3.13	0.67	0.69	1.38	0.68
SSR_80	53	4	2.72	0.70	0.64	1.10	0.63
SSR_89	49	6	2.04	0.53	0.52	1.01	0.51
SSR_95	54	8	4.23	0.87	0.77	1.62	0.76
Mean	51	5.48	2.95	0.60	0.62	1.20	0.62
St. dev.	4	1.93	1.06	0.24	0.14	0.37	0.14

2.4 Investigating Mandevilla germplasm by intersecting SSR data, ploidy level and DNA barcoding

The newly developed SSR set was used to investigate the level of genetic differentiation existing among the 55 accessions by calculating genetic similarity estimates for all possible pairwise comparisons. Genetic similarity (GS) values (mean = 0.72) ranged from 0.51 (2016 vs. 2009) to 1.00 (2001 vs. 2002, 2001 vs. 2008 and 2002 vs. 2008; **Supplementary Table 5**). Very low similarity values (such as in the case of 2016 vs. 2009) could be considered an indication that they actually belong to different species. In contrast, similarity scores close to 1.00 were often confirmed at the phenotypic level. For example, 2001, 2002 and 2008 were morphologically very similar. The common origin of these three samples is also confirmed by the fact that they were all constituted by the same breeder. Furthermore, the 2003 sample, which is known to be the result of a 2001 spot mutation, shows with the latter a degree of genetic similarity close to 100% (0.99). The GS-based UPGMA tree revealed a marked grouping of the samples into six clusters, highlighted in **Figure 4, panel A** with different colors.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0255.v1

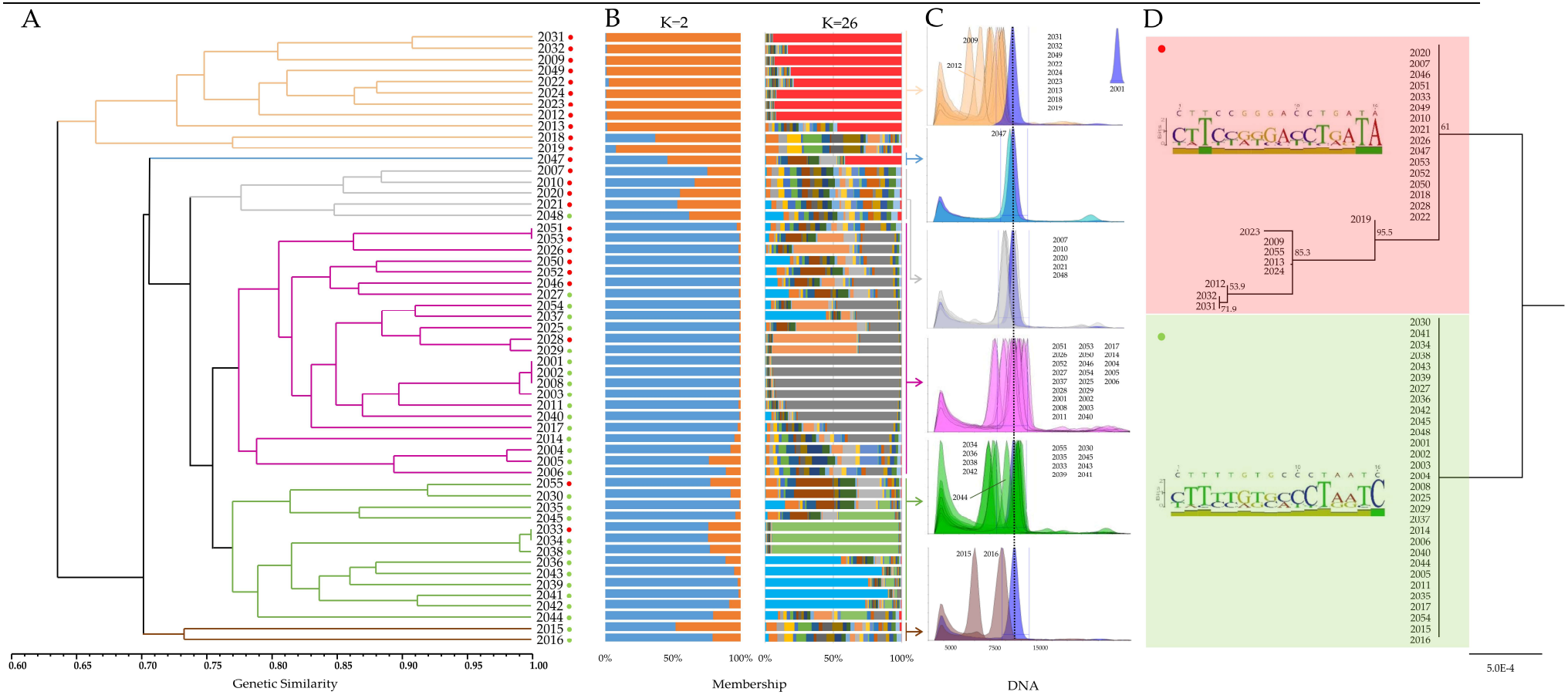


Figure 4 (A) UPGMA dendrogram based on a pairwise genetic similarity matrix (Supplementary Table 5). Six main clusters highlighted with different colors (ocher, light blue, grey, violet, green and brown) were found within the germplasm collection of Mandevilla (n=55). Red and green dots indicate the membership of each sample to one of the two clusters identified by targeted-region DNA sequencing (Figure 4D). **(B)** Genetic structure of the Mandevilla germplasm collection as estimated by STRUCTURE using the SSR marker dataset. Each sample is represented by a vertical histogram partitioned into K=2- and K=26-colored segments that represent the estimated membership (Supplementary Figure 1). The proportion of ancestry (%) is reported on the abscissa axis. **(C)** DAPI-based flow cytometry measurements of the 55 Mandevilla accessions. Each sample was analyzed by co-staining its nuclei along with those extracted from 2001 (the reference sample whose degree of ploidy was known, i.e., triploid). Samples were then grouped and analyzed according to the six clusters identified in Figure 4A. **(D)** Neighbor-joining tree built by concatenating the ITS1 and *rbc*L sequences of each sample and aligning them through the Clustal omega algorithm. Bootstrap values are reported. For each of the two groups identified through DNA barcoding (*rbc*L) and DNA haplotyping (ITS1), in the left part of the panel, a polymorphic site-based sequence logo is reported.

 doi:10.20944/preprints202204.0255.v1

Microsatellite data were also used to investigate the genetic structure of the collection and all the possible ancestors. Following the procedure described by Evanno et al. [32], a clear maximum for ΔK values was found at K = 26 and K=2 (**Supplementary Figure 1**). For K=2 (**Figure 4**, panel B), all individuals grouping together in the UPGMA tree under the ocher cluster showed an individual membership to their founding group (orange) higher than 92% (except for 2018, 63%). In contrast, the 23 samples from the largest UPGMA cluster (violet) exhibited an average membership to the other founding group (blue) higher than 97%. All other samples were, with few exceptions (i.e., 2035, 2039, 2041 and 2045), admixed. Analyzing the same collection for K = 26 (**Figure 4**, panel B), some subgroups emerged, both confirming the groupings identified for K = 2 and strengthening the common origin of some accessions. This is the case for 2001, 2002, 2003 and 2008 (gray group) or 2036, 2039, 2041, 2042 and 2043 (light blue group). Additionally, 2033, 2034 and 2038 proved to share a common ancestor (pale green group), as already demonstrated by their average genetic similarity (99%) and by the information available from the literature. An attempt to correlate this genetic evidence with the DNA content of each sample was made by taking advantage of the cytometric analyses.

The 55 Mandevilla accessions were therefore analyzed by costaining their nuclei (DAPI) along with those extracted from 2001, as this latter was the only sample whose degree of ploidy was known (triploid). Unfortunately, considering that each accession is, in all likelihood, the result of repeated interspecific crossings, the use of 2001 as a reference to establish the degree of ploidy of the other 54 samples proved unsuccessful. In fact, the ratio between the median fluorescence intensity of each sample (MFI_s) and the MFI of 2001 (MFI_{2001}) oscillated between 0.49 and 1.14 (**Supplementary Table 6**). For this reason, except for those accessions whose MFI_s/MFI_{2001} ratio was approximately 1 (and for which triploidy can be postulated), it was not possible to assign a defined ploidy value to most of the samples. For example, in the case of 2009 and 2015, with MFI_s/MFI_{2001} ratios of approximately 0.5, it is impossible to explain a noninteger ploidy value (i.e., 1.5), and it is more likely that these samples represent species or interspecific hybrids different from those of 2001. The genus *Mandevilla* is characterized by a variable number of chromosomes (i.e., $2n$=16, $2n$=20, $2n$=22 [33,34]) and without karyological data, defining the number of chromosomes and chromosome pairs inherited from an interspecific hybrid is challenging. At the same time, the strong interfertility existing among the species of this genus could have favored the occurrence of dysploidy [33], complicating the karyotypic scenario through gains and losses of single chromosomes or fission and/or fusion of chromosome segments [35]. Based on this new perspective, the effectiveness of the SSR set in analyzing all species/interspecific hybrids of the germplasm must be acknowledged. The fact that all 23 loci showed at most two alleles per locus would also support the idea that the genomic sequences used for SSR selection were derived from a diploid progenitor common to all accessions.

Although the lack of adequate references and genealogical background hindered the definition of the ploidy level, the cytometry data were still considered to highlight the extent—in terms of DNA content—of the gap existing between each sample and 2001 (**Figure 4**, panel C). In this regard, it is worth highlighting how the lower values of DNA content were registered in the two sample groups (ocher and brown) that - based on the SSR-based UPGMA tree - clustered apart from 2001 (mean MFI_{ocher}/MFI_{2001} = 0.77 and mean MFI_{brown}/MFI_{2001} = 0.69; **Supplementary Table 6**). In contrast, the average MFI_{violet}/MFI_{2001} ratio calculated for samples sharing the same SSR-based cluster as 2001 (violet) was close to 1 (0.99). Although it is ventured to find an explanation capable of correlating DNA content and SSR-based genetic analyses, it could be hypothesized that sharing a common genealogy (e.g., sharing a parent/ancestor) could explain both a high genetic similarity and a comparable DNA content.

To clarify the genetic relationships theorized through the SSR data and the MFI_s/MFI_{2001} ratios, a target-DNA region sequencing analysis was conducted by means of a plastidial DNA barcoding (*rbc*L) and a nuclear DNA haplotyping (ITS1). Various coding and noncoding regions of plastid genomes and nuclear regions have been proposed, alone

 doi:10.20944/preprints202204.0255.v1

or in combination, for DNA barcoding studies in Apocynaceae. Among these spacers, the *trnH-psbA* spacer was proposed in combination with *matK* [36] or ITS2 [37], *rbc*L was combined with *atpB*, *rpoC1*, *trnH-psbA* and ITS1/2 [38], and the *trnL-F* efficiency was compared with that of *matK*, *rbc*L and *trnH-psbA* [39]. By reviewing the main DNA barcoding studies performed in this family, we found contrasting opinions on which region is the most recommended, as the type of analysis (intragenus vs. intergenera) seems to greatly influence the suitability of the barcoding region. However, several works recognized a higher efficiency in the combined use of plastidial and nuclear markers [37,38,40]. For this reason, we chose both the plastidial *rbc*L region - one of the two core barcodes established by the Consortium for Barcode of Life, (CBOL [41]) - and ITS1, a supplementary nuclear barcode candidate suggested again by the CBOL.

The obtained sequences were 632 bp (*rbc*L) and 306 bp (ITS1) long. In particular, *rbc*L was polymorphic only at position 575 (A>C), splitting the core collection into two groups of 30 and 25 accessions. Both *rbc*L sequences were used to interrogate the BOLD database, and the best match was *Mandevilla sanderi* (100% query coverage), with a slight difference in terms of identity values (100%, first sequence and 99.84% second sequence). It can be hypothesized that the maternal lineage of all the samples is likely to be the same and that the SSR panel worked in all the interspecific hybrids of the germplasm because they were (casually) selected on the portion of the maternal genome common to all. It is worth emphasizing that the smallest group (575: C) included all the samples belonging to the SSR-based ocher cluster (red dots in **Figure 4**, panel A), in line with what emerged from the genetic structure analysis and from the considerations made about the MFI_s/MFI_{2001} ratios. This would further support the hypothesis that these samples have a very distinct origin from the other clusters. ITS1, despite being shorter than *rbc*L, was polymorphic in 15 positions (15/306, **Supplementary Table 7**). All the *rbc*L sequences, searched in GenBank through blastN, had *Mandevilla atroviolacea* as the best match with full query coverage (always 100%) and an identity score ranging from 98.52% to 95.45%. This relatively low degree of genetic identity must be interpreted considering the very limited number of ITS1 sequences available in GenBank for the *Mandevilla* genus. It is extremely likely that the species/hybrids to which the germplasm samples belong are not represented in GenBank and thus cannot be properly identified. However, by predicting - at each polymorphic position - the genotypic combinations of the two putative parents, we were able to predict the specific parental genotypes that distinguished the two groups of samples previously identified by means of *rbc*L (**Supplementary Table 7**).

Finally, an NJ dendrogram was built by concatenating the ITS1 and *rbc*L sequences of each sample and aligning them through the Clustal omega algorithm. Overall, samples clustered in the same two groups identified solely based on *rbc*L (**Figure 4**, panel D). In addition, a further subgroup that included all the samples from the SSR-based ocher cluster (**Figure 4**, panel A) was identified, again corroborating the hypothesis of a distinct origin.

3. Materials and Methods

3.1 Plant materials

Fifty-five samples from the *Mandevilla* genus were collected as representatives of the current market. With few exceptions, no information was available about any possible origin, relatedness or ploidy level of the accessions. Moreover, many, if not all, accessions could be interspecific hybrids. Due to the economic interest behind this species, breeding companies often contain all this information. In light of these considerations, this study always refers to the genus *Mandevilla* without specifying any species.

3.2 Genome size estimate through flow citometry

The genome size of the "Mandevilla 2001" sample, subsequently used for genome sequencing, was determined through flow cytometry of propidium iodide (PI)-stained nuclei, following the procedure described by the CyStain PI Absolute P protocol (Sysmex Partec, Görlitz, Germany). One hundred milligrams of fresh leaf tissue was chopped with

 doi:10.20944/preprints202204.0255.v1

a razor blade along with 0.5 ml of Nuclei Extraction Buffer (Sysmex Partec), incubated for 45 minutes at room temperature and filtered using 30 µm CellTrics (Sysmex Partec). Two milliliters of staining solution (1982 µl of Staining Buffer, 12 µl of PI and 6 µl of RNAse A 3.3 ng/µl) was then added to each filtered sample, and the resulting solution was placed on ice in the dark for 45 minutes. Analyses were run by setting the following parameters: Nd-YAG green laser: $\lambda = 532$ nm; 30 mW, flow rate of 4 µl/s. *Raphanus sativus, Glycine max* and *Solanum lycopersicum* seeds with known 2C DNA content were kindly provided by Prof. Dolezel (https://olomouc.ueb.cas.cz/en/technology/flow-cytometry-1/reference-dna-standards), adopted as reference standards, and their relative fluorescence was used to estimate the genome size of the 2001 sample. Fluorescence histograms were evaluated using FCS Express 5 Flow software (Sysmex Partec), and c-values were inferred by comparing the sample and standard at G0/G1 peak positions.

3.3 DNA extraction, library preparation and sequencing

Due to its historical relevance, the "Mandevilla 2001" variety was selected for whole genome sequencing analysis. Leaves were collected, snap-frozen in liquid nitrogen upon harvesting and stored at −80 °C until further processing. Genomic DNA was isolated using the cetyltrimethylammonium bromide (CTAB) protocol [42]. gDNA integrity was first evaluated through an electrophoretic run (0.8% agarose/1× TAE gel containing 1× Sybr Safe DNA stain; Life Technologies, Carlsbad, CA, USA) and an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). Concentration and purity (260/280-nm and 260/230-nm absorbance ratios) were spectrophotometrically assessed thorough a NanoDrop 2000c spectrophotometer (Thermo Scientific, Waltham, MA, USA).

Library preparation was accomplished using ~500 ng of gDNA and the Illumina Nextera DNA Flex library preparation kit (Illumina, Inc., San Diego, CA, USA) following the protocol provided by the company. The library was sequenced on an Illumina NovaSeq system using paired-end, 100-bp-read chemistry and with an average insert size of 550 bp.

3.4 Plastidial genome assembly and phylogenetic analysis

NOVOPlasty 4.2 [43] was specifically used to assemble the chloroplast genome, adopting the cpDNA of *Rhazya stricta* (NC_024292.1, from NCBI) as the reference sequence and the *rbc*L gene of *M. sanderi* (X91764.1 from NCBI) as seed input. *R. stricta* was selected as the phylogenetically closest species to *Mandevilla* with both organelle genomes available. The following parameters were also specified: automatic insert size detection, a genome size range from 120,000 to 180,000 (based on an estimate derived from *Rhazya stricta)*, and a K-mer value of 33. The cp genome was annotated using GeSeq [44] using *Rhazya stricta* as a reference. OGDRAW [45] was finally employed for the graphical visualization of the circular DNA. To investigate the systematic relationships existing between the *Mandevilla* genus and the rest of the Apocynaceae family, a DNA superbarcoding-based analysis was performed by aligning the newly assembled cpDNA with the 33 plastid genomes available in NCBI [46] for the abovementioned family (**Supplementary Table 8**). The cpDNA of *Gentiana officinalis* was selected as the outgroup since its family (Gentianaceae) is from the same order (Gentianales) of Apocynaceae. The 35 plastomes were aligned using MAFFT v7.450 [47] under default parameters, while Geneious prime (Biomatters, Inc, San Diego, CA, USA) was used to reconstruct the phylogenetic relationships based on the neighbor-joining (NJ) method with 1000 bootstrap replicates.

3.5 Nuclear genome assembly and annotation

For nuclear genome assembly, raw sequences were first processed using fastp software [48] to remove the adapter sequences and to trim low-quality bases using the following parameters: qualified_quality_phred 20, unqualified_percent_limit 30, average_qual 25, low_complexity_filter = True, complexity_threshold 30, length_required 40. The filtered reads were then assembled using MaSuRCA [49] under the following conditions: extend_jump_reads = 0; graph_kmer_size = auto; use_linking_mates = 1; use_grid = 0; lhe_coverage = 25; mega_reads_one_pass = 1; ca_parameters = cgwerrorrate = 0.15;

 doi:10.20944/preprints202204.0255.v1

close_gaps = 1; jf_size = 60000000000. The same software was also used to predict a k-mer-based haploid genome size estimate by evaluating K-values ranging from 19 to 31.

According to the estimated haploid genome size, raw data were used to compute the sequencing depth by means of the Lander/Waterman equation:

$$C = LN/G \qquad (1)$$

where C is the average coverage, L is read length, N is the number of reads and G is the haploid genome length.

Assembly statistics were calculated using the NGS QC Toolkit v.2.3.3 [50], while a BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis [51], based on 2,326 single-copy orthologs included in the eudicots_odb10 database [52], was performed to provide a quantitative assessment of the completeness in terms of the expected gene content of the genome assembly.

The newly assembled genome was annotated against the *Coffea canephora* (Rubiaceae family, Gentianales order) proteome (GCA_900059795.1, 25,574 proteins), with this species being the phylogenetically closest to the *Mandevilla* genus with a well-assembled and annotated genome. A tBLASTn-based approach (E-value ≤1e-10, BLAST+ v.2.3.0) was employed by using the proteome as a query to interrogate the assembled genome, while ShinyGO v.0.741 was used for Gene Ontology enrichment analysis [53].

As monoterpene indole alkaloids (MIAs) are a large group of plant compounds with enormous pharmaceutical potential that have been isolated mainly from species belonging to the Apocynaceae family [22], we tried to decipher whether Mandevilla has the orthologous enzymes necessary to produce these molecules. At this aim, 10 amino acid sequences associated with enzymes involved in the biosynthesis of the three main MIAs (i.e., catharanthine, tabersonine and vincadifformine [23,25]) in *Catharanthus roseus* (Apocynaceae) were retrieved from NCBI and used as queries in a tBLASTn-based approach (E-value ≤ 1e-40) against the genome.

3.6 Simple sequence repeat (SSR) identification, validation, and data analysis

Microsatellite regions (or SSRs) were searched through the MIcroSAtellite (MISA) Identification Tool Perl script [54], screening the genome for mono-, tri-, tetra-, penta- and hexanucleotide repeat motifs with minimum repeat numbers of 20, 10, 7, 5, 5, and 5, respectively. The maximum number of nucleotides interrupting two SSR regions in a compound microsatellite was set at 200 bp, and the space between imperfect SSR stretches was set at 5 bp. Finally, a total of 100 primer pairs (**Supplementary Table 4**) were randomly designed using Geneious Prime 2020 software (Biomatters Ltd.), according to the following considerations: (i) dinucleotide or trinucleotide repeat motif, (ii) length of the repeat motif ≥20 times (i.e., ten dinucleotide repetitions or seven trinucleotide repetitions) and (iii) melting temperature (T_m) always between 55 °C and 57 °C to facilitate their use in multiplex reactions. For this latter reason, the 5′ end of each forward primer was tagged with an oligonucleotide tail (M13, PAN1, PAN2 or PAN3) to be employed in PCRs in combination with a complementary fluorophore-labeled (6-FAM, VIC, NED and PET were used as fluorophores) oligonucleotide. This three-primer-based strategy is a modified version of what was reported by Schuelke [55] and originally described in [56].

The amplification efficiency and the polymorphic rate of the 100 SSR primer pairs were preliminarily tested in single and multiple reactions on 4 samples (including 2001), selected on the basis of marked phenotypic diversity and whose DNA was isolated using the DNeasy Plant Pro Kit (Qiagen, Hilden, Germany). The 23 SSR marker loci that amplified efficiently generating unambiguous and polymorphic profiles (**Table 2**) were then organized in 4 multiplex and validated on 55 samples belonging to a Mandevilla germplasm whose gDNA was extracted as previously described.

PCRs were performed in a total volume of 20 µl containing approximately 20 ng of gDNA template, 1× Platinum Multiplex PCR Master Mix (Applied Biosystems, Carlsbad, CA, USA), GC enhancer 10% (Applied Biosystems), 0.05 µM tailed forward primer (Invitrogen, Carlsbad, CA, USA), 0.1 µM reverse primer (Invitrogen), 0.23 µM universal primer (Invitrogen) and sterile water to volume. A 9600 Thermal Cycler (Applied Biosystems) with 96-well plates was used for PCRs setting the following thermal conditions: 5 min at 95 °C, followed by three cycles at 95 °C for 30 s and at 54 °C for 45 s, which decreased by 1 °C with each cycle, and at 72 °C for 45 s; then 37 cycles at 95 °C for 30 s, at 51 °C for 45 s, and at 72 °C for 45 s. Reactions were terminated with a final extension of 30 min at 60 °C.

PCR products were first run by means of agarose gel electrophoresis (agarose 2% agarose/1× TAE gel containing 1× Sybr Safe DNA stain; Life Technologies) and visualized on an Uvidoc HD6 transilluminator (Uvitec, Cambridge, UK) equipped with a digital camera. They were then dried at 65 °C for 1 h and subjected to capillary electrophoresis (ABI PRISM 3130xl Genetic Analyzer, Thermo Fisher) and LIZ500 (Applied Biosystems) as the molecular weight standard.

The allele size of each SSR locus was determined through Peak Scanner Software 1.0 (Applied Biosystems), while the observed (Ho) and expected (He) homozygosity, the Shannon index (I) of phenotypic diversity, and the number of observed (Na) and effective (Ne) alleles were calculated with POPGENE v.1.32 software [57]. The polymorphism information content (PIC) and thus the informativeness of each microsatellite were assessed with the Excel Microsatellite Toolkit [58] as:

$$PIC=1-\sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2 \qquad (2)$$

Pairwise genetic similarity estimates were calculated in all possible comparisons among the 55 Mandevilla samples using NTSYS v.2.2 software [59] and by applying the Rohlf coefficient. The resulting triangular matrix was used, in turn, to build a UPGMA dendrogram. Finally, an attempt to reconstruct the genetic structure of the core collection was accomplished by modeling a Bayesian clustering algorithm available in STRUCTURE v.2.2 [60]. Ten replicate simulations were conducted for each value of K (range: 2-30) as suggested by Evanno et al. [32], while a burn-in of $2 \cdot 10^5$ and a final run of 10^6 Markov chain Monte Carlo (MCMC) steps were set. STRUCTURE HARVESTER [61] was run to estimate the most likely K value, while estimates of membership were plotted as histograms using an Excel spreadsheet.

3.7 Ploidy level estimates through flow citometry

The ploidy level of the 55 accessions (each analyzed in three biological replicates) was studied through flow cytometry (CyFlow Ploidy Analyzer) of 4',6-diamidino-2-phenylindole (DAPI)-stained nuclei following the procedure described by the CyStain UV Precise P protocol (Sysmex Partec). One hundred milligrams of fresh leaf tissue taken from each of the three biological replicates were co-chopped with a razor blade in a Petri dish with 0.5 ml of Nuclei Extraction Buffer (Sysmex Partec) and incubated for 45 minutes at room temperature. After filtering (30 µm of CellTrics®, Sysmex Partec), 2 ml of staining buffer was added to each sample and incubated for 60 s before analysis (Nd-YAG green laser: $\lambda = 532$ nm; 30 mW, flow rate of 4 µl/s). Fluorescence histograms were evaluated using FCS Express 5 Flow software (Sysmex Partec), and ploidy levels were inferred by comparing each sample with the ploidy level of the 2001 sample.

3.8 Assessing the phylogenetic origin of the Mandevilla germplasm

The DNA of the 55 samples previously genotyped with SSR markers and cytometrically analyzed for ploidy level investigation was also analyzed at the plastidial gene encoding the ribulose bisphosphate carboxylase large chain (*rbc*L) and at nuclear internal transcribed spacer 1 (ITS1). Primer pairs adopted, along with the relative nucleotide sequences, are available in Scariolo et al. [62]. PCRs were performed in 25 µL containing 50 ng of genomic DNA as a template, 12.5 µL of MangoMix (Bioline, London, UK), 2 µL of each primer (10 mM) and sterile water to reach the final volume. A Veriti 96-Well Thermal

 doi:10.20944/preprints202204.0255.v1

Cycler (Applied Biosystems) was used to carry out the amplifications by setting the following parameters: 2 min at 95 °C; 35 cycles at 95 °C for 30 s, 55 °C for 45 s, and 72 °C for 45 s; and a final extension at 72 °C for 10 min. PCR products were purified with ExoSAP-IT PCR Product Cleanup Reagent (Thermo Fisher) and sequenced on an ABI 3730XL Genetic Analyzer (Applied Biosystems). Chromatograms were evaluated using Geneious Prime software (Geneious software), and sequences were trimmed at the 5′ and 3′ positions to remove the low-quality regions. The clean *rbc*L and ITS1 sequences of each sample were examined in the BOLD system and GenBank, respectively, by means of a BLASTn search and were then concatenated and globally aligned with those of the other samples (Clustal omega algorithm, Geneious software). The resulting multiple alignment was used for the construction of a neighbor-joining (NJ) tree using the Juke-Cantor algorithm (a bootstrap analysis was conducted to measure the stability of the computed branches with 1,000 resampling replicates). Polymorphic sites were used to create a logo graph.

4. Conclusions

In this study, we attempted to lay the foundations for the genomic and cytometric characterization of the genus *Mandevilla*. However, the road is still indisputably long. We managed to assemble the cpDNA (of great help for phylogenetic studies) and to produce a first draft of the nuclear genome, but the use of long reads-based sequencing platforms will be decisive in the future for improving the assembly (down to the chromosomal level) and for reconstructing mitochondrial DNA (mtDNA). The exclusive use of short read-based sequencing platforms (i.e., Illumina) makes the assembly of mtDNA extremely challenging and characterized by large repeated regions, nuclear and plastid-deriving sequences and several possible configurations (linear, circular or branched) [28]. The genomic contigs were useful to identify thousands of genes, some of which suggested the possibility that Mandevilla could produce MIAs of pharmaceutical interest, similar to many other species of the same family. Further chromatographic analyses will be crucial to elucidate if and possibly which MIA Mandevilla is able to synthetize. The genome draft was also of great help for the development of a robust panel of SSR markers, which have recently started to be used to assist breeding programs in noncrop species, including ornamental plants. As far as cytometry is concerned, the situation turned out to be, as expected, very complex to decipher. Although the intersection of cytometry data and k-mer estimates allowed us to infer the triploidy of the reference sample used for genome sequencing, the ploidy level for the different accessions of the germplasm still remains to be defined. In fact, by comparing their FMI with the FMI values of the reference, noninteger values were obtained. This is explainable by interspecific hybridization events (partially confirmed through DNA barcoding data), which led to dysploidy phenomena, gains and losses of single chromosomes or fission and/or fusion of chromosome segments. In this context, karyological analyses coupled with genomic *in situ* hybridization (GISH) studies could represent a great deal of help in assessing the origin, divergence, relationships and evolution of these hybrids. At the same time, even the enhancement of DNA barcoding databases could solve some uncertainties related to the possible parents of the aforementioned hybrids. In our study, assuming that *M. sanderi* was always used as maternal species, it remains to be clarified, due to lack of information, which species were used as paternal parents in the different crossing and/or backcrossing cycles of the breeding programs.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, **Figure S1**: Definition of the number of ancestral Mandevilla populations based on the SSR marker dataset. ΔK values are represented by the blue line, while the blue points indicate the mean LnP(D) ± SD values. Mean LnP(D) ± SD (over 10 runs) is a function of K, as L′(K) = ΔLnP(D) and mean ΔK is calculated as |L″(K)|/(SD(L(K)). **Table S1:** Genes of Mandevilla chloroplast genome. [1]indicates introns-containing genes. **Table S2:** tBLASTn results obtained by using the proteome of *Coffea canephora* (GCA_900059795.1) as a query to interrogate the assembled genome (E-value ≤1e-10, BLAST+ v.2.3.0). **Table S3**: Results obtained by using 10 amino acid sequences associated to enzymes in-

volved in the biosynthesis of the three main MIAs (i.e. catharanthine, tabersonine and vincadifformine) in *Catharanthus roseus* (Apocynaceae) as query in a tBLASTn-based approach (E-value ≤ 1e-40) against the assembled genome of Mandevilla. **Table S4:** Primer pairs randomly designed on the newly assembled genome of Mandevilla and able to amplify as many SSR loci. For each SSR locus, SSR name, SSR-containing contig, location of each SSR-containing contig mapped onto the *Coffea arabica* genome, primer sequences, mean SSR locus size, melting temperatures, anchor used and SSR motif are reported. **Table S5:** Genetic similarity matrix based on 23 SSR. In the first column, the observed heterozygosis is indicated too. **Table S6**: Cytometric analysis results for 55 Mandevilla accessions analyzed using 2001 sample as reference. **Table S7:** ITS1-based haplotypes reconstructed for the 55 Mandevilla accessions. Samples are organized in two subgroups based on the *rbc*L sequence alignment: pink samples shared a "C" in position 575, blue samples shared a "A" in the same position. Predicted parental genotypes for each polymorphic positions are reported. Parental genotypes discriminating the two samples groups are highlighted in red bold. **Table S8:** Plastomes used for phylogenetic analysis within the Apocynaceae family. Latin name, taxonomic family, BioProject, size (in Mb) of each plastome, GC % content and GenBank IDs are indicated for each species included in the analysis

Author Contributions: Conceptualization, F.P. and G.B.; methodology, S.D, F.S. and F.P.; formal analysis, S.D. F.S.; data analysis, F.P., S.D. and F.S.; writing—original draft preparation, F.P.; writing—review and editing, F.P., F.S., S.D., G.B.S., M.G. and G.B.; supervision, G.B.; project administration, G.B.; funding acquisition, G.B. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Gruppo Padana Ortofloricoltura company (Paese, TV, Italy), within the research contract signed with the Department of Agronomy, Food, Natural resources, Animals and Environment (DAFNAE), University of Padua, Italy.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Genbank under accession IDs OM489306 (cpDNA) and PRJNA802340 (nDNA)

Acknowledgments: We convey our thanks to Prof. Jaroslav Doležel of the Institute of Experimental Botany (Olomouc, Czech Republic) for providing the seeds reference for cytometry analyses. We would like to thank the graduated fellow Elisa Pasquali and the undergraduate student Chandana Mulagala of the University of Padova for their help in the ploidy and SSR analysis. Finally, we thank Gruppo Padana Ortofloricoltura company for the scientific collaboration and technical assistance in all the activities of this project.

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

References

1. Eltlbany, N.; Prokscha, Z.Z.; Castañeda-Ojeda, M.P.; Krögerrecklenfort, E.; Heuer, H.; Wohank, W.; Ramos, C.; Smalla, K. A new bacterial disease on Mandevilla sanderi, caused by Pseudomonas savastanoi: Lessons learned for bacterial diversity studies. *Appl. Environ. Microbiol.* **2012**, *78*, 8492–8497, doi:10.1128/AEM.02049-12.

2. Bhadane, B.S.; Patil, M.P.; Maheshwari, V.L.; Patil, R.H. Ethnopharmacology, phytochemistry, and biotechnological advances of family Apocynaceae: A review. *Phyther. Res.* **2018**, *32*, 1181–1210, doi:10.1002/ptr.6066.

3. Naidoo, C.M.; Naidoo, Y.; Dewir, Y.H.; Murthy, H.N.; El-Hendawy, S.; Al-Suhaibani, N. Major bioactive alkaloids and biological activities of tabernaemontana species (Apocynaceae). *Plants* **2021**, *10*, 313, doi:10.3390/plants10020313.

4. Wen, S.; Chen, Y.; Lu, Y.; Wang, Y.; Ding, L.; Jiang, M. Cardenolides from the Apocynaceae family and their anticancer activity. *Fitoterapia* **2016**, *112*, 74–84, doi:10.1016/j.fitote.2016.04.023.

5. Anand, U.; Nandy, S.; Mundhra, A.; Das, N.; Pandey, D.K.; Dey, A. A review on antimicrobial botanicals, phytochemicals and natural resistance modifying agents from Apocynaceae family: Possible therapeutic approaches against multidrug resistance in pathogenic microorganisms. *Drug Resist. Updat.* **2020**, *51*, 100695, doi:10.1016/j.drup.2020.100695.

6. Oder, A.; Lannes, R.; Viruel, M.A. A set of 20 new SSR markers developed and evaluated in mandevilla Lindl. *Molecules* **2016**, *21*, 1316, doi:10.3390/molecules21101316.

7. de Candolle, A. *Prodromus systematis naturalis regni vegetabilis, sive, Enumeratio contracta ordinum generum specierumque plantarum huc usque cognitarium, juxta methodi naturalis, normas digesta.*; Parisii: Sumptibus Sociorum Treuttel et Würtz, 1824-73., 1844;

8. Woodson Jr, R.E. Studies in the Apocynaceae . IV . The American Genera of Echitoideae. *Ann. Missouri Bot. Gard.* **1933**, *20*, 605–790, doi:https://doi.org/10.2307/2394198.

9. Flowers, S. Suntory flowers varieties Available online: https://suntoryflowers.com/varieties/ (accessed on Oct 31, 2021).

10. Royal Botanical Garden Kew Mandevilla Lindl.

11. Simões, A.O.; Endress, M.E.; Van Der Niet, T.; Kinoshita, L.S.; Conti, E. Is Mandevilla (Apocynaceae, Mesechiteae) monophyletic? Evidence from five plastid DNA loci and morphology. *Ann. Missouri Bot. Gard.* **2006**, *93*, 565–591, doi:10.3417/0026-6493(2006)93[565:IMAMME]2.0.CO;2.

12. Zhang, L.H.; Byrne, D.H.; Ballard, R.E.; Rajapakse, S. Microsatellite marker development in rose and its application in tetraploid mapping. *J. Am. Soc. Hortic. Sci.* **2006**, *131*, 380–387, doi:10.21273/jashs.131.3.380.

13. Lee, S. Il; Park, K.C.; Song, Y.S.; Son, J.H.; Kwon, S.J.; Na, J.K.; Kim, J.H.; Kim, N.S. Development of expressed sequence tag derived-simple sequence repeats in the genus Lilium. *Genes and Genomics* **2011**, *33*, 727–733, doi:10.1007/s13258-011-0203-1.

14. Pourkhaloee, A.; Khosh-Khui, M.; Arens, P.; Salehi, H.; Razi, H.; Niazi, A.; Afsharifar, A.; van Tuyl, J. Molecular analysis of genetic diversity, population structure, and phylogeny of wild and cultivated tulips (Tulipa L.) by genic microsatellites. *Hortic. Environ. Biotechnol.* **2018**, *59*, 875–888, doi:10.1007/s13580-018-0055-6.

15. Jo, K.M.; Jo, Y.; Chu, H.; Lian, S.; Cho, W.K. Development of EST-derived SSR markers using next-generation sequencing to reveal the genetic diversity of 50 chrysanthemum cultivars. *Biochem. Syst. Ecol.* **2015**, *60*, 37–45, doi:10.1016/j.bse.2015.03.002.

16. Royal Botanical Garden Kew Plant DNA C-value database Available online: https://cvalues.science.kew.org/search/angiosperm (accessed on Oct 31, 2021).

17. Bennett, M.D.; Price, H.J.; Johnston, J.S. Anthocyanin inhibits propidium iodide DNA fluorescence in Euphorbia pulcherrima: Implications for genome size variation and flow cytometry. *Ann. Bot.* **2008**, *101*, 777–790, doi:10.1093/aob/mcm303.

18. Dolezel, J.; Bartos, J.; Voglmayr, H.; Greilhuber, J. Nuclear DNA content and genome size of trout and human.

Cytom. Part A **2003**, *51A*, 127–128, doi:10.1002/cyto.a.10013.

19. Plunkett, G.; Downie, S.R. Expansion and Contraction of the Chloroplast Inverted Repeat in Apiaceae Subfamily Apioideae. *Syst. Bot.* **2000**, *25*, 648–667.

20. Livshultz, T. The phylogenetic position of milkweeds (Apocynaceae subfamilies Secamonoideae and Asclepiadoideae): Evidence from the nucleus and chloroplast. *Taxon* **2010**, *59*, 1016–1030, doi:10.1002/tax.594003.

21. Livshultz, T.; Middleton, D.J.; Endress, M.E.; Justin, K.; Apocynaceae, S.L.; Livshultz, T.; Middleton, D.J.; Mary, E.; Williams, J.K. Phylogeny of Apocynoideae and the APSA Clade (Apocynaceae s. l.). *Ann. Missouri Bot. Gard.* **2007**, *94*, 324–359.

22. Balsevich, J. Monoterpene Indole Alkaloids from Apocynaceae other than Catharanthus roseus. In *Phytochemicals in Plant Cell Cultures*; Constabel, F., Vasil, I.K., Eds.; Academic Press, Inc: New York, USA, 1988; Vol. 5, pp. 371–384.

23. Qu, Y.; Safonova, O.; De Luca, V. Completion of the canonical pathway for assembly of anticancer drugs vincristine/vinblastine in Catharanthus roseus. *Plant J.* **2019**, *97*, 257–266, doi:10.1111/tpj.14111.

24. Qu, Y.; Thamm, A.M.K.; Czerwinski, M.; Masada, S.; Kim, K.H.; Jones, G.; Liang, P.; De Luca, V. Geissoschizine synthase controls flux in the formation of monoterpenoid indole alkaloids in a Catharanthus roseus mutant. *Planta* **2018**, *247*, 625–634, doi:10.1007/s00425-017-2812-7.

25. Qu, Y.; Easson, M.E.A.M.; Simionescu, R.; Hajicek, J.; Thamm, A.M.K.; Salim, V.; De Luca, V. Solution of the multistep pathway for assembly of corynanthean, strychnos, iboga, and aspidosperma monoterpenoid indole alkaloids from 19E-geissoschizine. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 3180–3185, doi:10.1073/pnas.1719979115.

26. Mohammed, A.E.; Abdul-Hameed, Z.H.; Alotaibi, M.O.; Bawakid, N.O.; Sobahi, T.R.; Abdel-Lateff, A.; Alarif, W.M. Chemical diversity and bioactivities of monoterpene indole alkaloids (MIAs) from six Apocynaceae genera. *Molecules* **2021**, *26*, 488, doi:10.3390/molecules26020488.

27. Palumbo, F.; Vannozzi, A.; Magon, G.; Lucchin, M.; Barcaccia, G. Genomics of flower identity in grapevine (Vitis vinifera L.). *Front. Plant Sci.* **2019**, *10*, 316, doi:10.3389/fpls.2019.00316.

28. Palumbo, F.; Vannozzi, A.; Barcaccia, G. Impact of genomic and transcriptomic resources on apiaceae crop breeding strategies. *Int. J. Mol. Sci.* **2021**, *22*, 9713, doi:10.3390/ijms22189713.

29. Carvalho, A.B.; Dupim, E.G.; Goldstein, G. Improved assembly of noisy long reads by k-mer validation. *Genome Res.* **2016**, *26*, 1710–1720, doi:10.1101/gr.209247.116.

30. Srivastava, S.; Avvaru, A.K.; Sowpati, D.T.; Mishra, R.K. Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics* **2019**, *20*, 153, doi:10.1186/s12864-019-5516-5.

31. Botstein, D.; White, R.L.; Skolnick, M.; Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **1980**, *32*, 314–331.

32. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **2005**, *14*, 2611–20.

33. Andriyanova, E.A.; Lomonosova, M.N.; Berkutenko, A.N. IAPT/IOPB chromosome data 17. *Taxon* **2014**, *63*, 1148–1155.

34. Sanso, A.M.; Xifreda, C.C. Karyotypes of Macrosiphonia petraea and M. virescens (Apocynaceae). *Bol. Soc. Argent. Bot.* **2000**, *35*, 291–295.

35. Winterfeld, G.; Ley, A.; Hoffmann, M.H.; Paule, J.; Röser, M. Dysploidy and polyploidy trigger strong variation of chromosome numbers in the prayer-plant family (Marantaceae). *Plant Syst. Evol.* **2020**, *306*, 36, doi:10.1007/s00606-020-01663-x.

36. Mahadani, P.; Sharma, G.D.; Ghosh, S.K. Identification of ethnomedicinal plants (Rauvolfioideae: Apocynaceae) through DNA barcoding from northeast India. *Pharmacogn. Mag.* **2013**, *9*, 255–263, doi:10.4103/0973-1296.113284.

37. Lv, Y.N.; Yang, C.Y.; Shi, L.C.; Zhang, Z.L.; Xu, A.S.; Zhang, L.X.; Li, X.L.; Li, H.T. Identification of medicinal plants within the Apocynaceae family using ITS2 and psbA-trnH barcodes. *Chin. J. Nat. Med.* **2020**, *18*, 594–605, doi:10.1016/S1875-5364(20)30071-6.

38. Selvaraj, D.; Sarma, R.K.; Shanmughanandhan, D.; Srinivasan, R.; Ramalingam, S. Evaluation of DNA barcode candidates for the discrimination of the large plant family Apocynaceae. *Plant Syst. Evol.* **2015**, *301*, 1263–1273, doi:10.1007/s00606-014-1149-y.

39. Cabelin, V.L.D.; Alejandro, G.J.D. Efficiency of matK, rbcL, trnH-psbA, and trnL-F (cpDNA) to molecularly authenticate Philippine ethnomedicinal Apocynaceae through DNA barcoding. *Pharmacogn. Mag.* **2016**, *12*, S384–S388, doi:10.4103/0973-1296.185780.

40. Mishra, P.; Kumar, A.; Sivaraman, G.; Shukla, A.K.; Kaliamoorthy, R.; Slater, A.; Velusamy, S. Character-based DNA barcoding for authentication and conservation of IUCN Red listed threatened species of genus Decalepis (Apocynaceae). *Sci. Rep.* **2017**, *7*, 14910, doi:10.1038/s41598-017-14887-8.

41. CBOL Plant Working Group A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12794–7.

42. Doyle, J.J.; Doyle, J.L. A rapid DNA isolation for small quantities of fresh leaf tissue. *Phytochem. Bull.* 1987, *19*, 11–15.

43. Dierckxsens, N.; Mardulyn, P.; Smits, G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **2017**, *45*, e18, doi:10.1093/nar/gkw955.

44. 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, doi:10.1093/nar/gkx391.

45. Greiner, S.; Lehwark, P.; Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **2019**, *47*, W59–W64, doi:10.1093/nar/gkz238.

46. U.S. National Library of Medicine Genome information by organisms Available online:

https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/apocynaceae (accessed on Oct 31, 2021).

47. Katoh, K.; Standley, D.M. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.

48. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890, doi:10.1093/bioinformatics/bty560.

49. Zimin, A. V.; Marçais, G.; Puiu, D.; Roberts, M.; Salzberg, S.L.; Yorke, J.A. The MaSuRCA genome assembler. *Bioinformatics* **2013**, *29*, 2669–2677, doi:10.1093/bioinformatics/btt476.

50. Patel, R.K.; Jain, M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One* **2012**, *7*, e30619, doi:10.1371/journal.pone.0030619.

51. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212, doi:10.1093/bioinformatics/btv351.

52. Afgan, E.; Baker, D.; Batut, B.; Van Den Beek, M.; Bouvier, D.; Ech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W537–W544, doi:10.1093/nar/gky379.

53. Ge, S.X.; Jung, D. ShinyGO: a graphical enrichment tool for animals and plants. *Bioin* **2018**, *36*, 2628–2629, doi:10.1101/315150.

54. Thiel, T.; Michalek, W.; Varshney, R.K.; Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor. Appl. Genet.* **2003**, *106*, 411–422, doi:10.1007/s00122-002-1031-0.

55. Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **2000**, *18*, 233–234, doi:10.1038/72708.

56. Palumbo, F.; Galla, G.; Martínez-Bello, L.; Barcaccia, G. Venetian local corn (Zea mays L.) germplasm: Disclosing the genetic anatomy of old landraces suited for typical cornmeal mush production. *Diversity* **2017**, *9*, 32, doi:10.3390/d9030032.

57. Yeh, F.C.; Yang, R.C.; Boyle, T.B.J.; Ye, Z.H.; Mao, J.. POPGENE, the user friendly shareware for population genetic analysis 1997.

58. Park, S.D.E. The Excel microsatellite toolkit 2001.

59. Rohlf, F.J. NTSYS-PC Numerical taxonomy and multivariate analysis system 1988.

60. Falush, D.; Stephens, M.; Pritchard, K.J. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **2003**, *164*, 1567–1587.

61. Earl, D.A.; VonHoldt, B.M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE

output and implementing the Evanno method. *Conserv. Genet. Resour.* **2012**, *4*, 359–361, doi:10.1007/s12686-011-9548-7.

62. Scariolo, F.; Palumbo, F.; Vannozzi, A.; Sacilotto, G.B.; Gazzola, M.; Barcaccia, G. Genotyping analysis by RAD-seq reads is useful to assess the genetic identity and relationships of breeding lines in Lavender species aimed at managing plant variety protection. *Genes (Basel).* **2021**, *12*, 1656, doi:10.3390/genes12111656.