Short Title: Comparison of *Nicotiana* plastid genomes

Plastid genomics of *Nicotiana* (Solanaceae): insights into molecular evolution, positive selection and the origin of the maternal genome of Aztec tobacco (*Nicotiana rustica*)

Furrukh Mehmood[1], Abdullah[1], Zartasha Ubaid[1], Iram Shahzadi[1], Ibrar Ahmed[2], Mohammad Tahir Waheed[1], Péter Poczai*[3], Bushra Mirza*[1]

[1]Department of Biochemistry, Quaid-i-Azam University, Islamabad, Pakistan

[2]Alpha Genomics Private Limited, Islamabad, Pakistan

[3]Finnish Museum of Natural History (Botany Unit), University of Helsinki, Helsinki, Finland

*Corresponding authors: **Bushra Mirza (bushramirza@qau.edu.pk)**

Péter Poczai (peter.poczai@helsinki.fi)

Abstract

The genus *Nicotiana* of the family Solanaceae, commonly referred to as tobacco plants, are a group cultivated as garden ornamentals. Besides their use in the worldwide production of tobacco leaves, they are also used as evolutionary model systems due to their complex development history, which is tangled by polyploidy and hybridization. Here, we assembled the plastid genomes of five tobacco species, namely *N. knightiana, N. rustica, N. paniculata, N. obtusifolia and N. glauca*. *De novo* assembled tobacco plastid genomes showed typical quadripartite structure, consisting of a pair of inverted repeats (IR) regions (25,323–25,369 bp each) separated by a large single copy (LSC) region (86,510 –86,716 bp) and a small single copy (SSC) region (18,441–18,555 bp). Comparative analyses of *Nicotiana* plastid genomes showed similar GC content, gene content, codon usage, simple sequence repeats, oligonucleotide repeats, RNA editing sites and substitutions with currently available Solanaceae genomes sequences. We identified twenty highly polymorphic regions mostly belonging to intergenic spacer regions (IGS), which could be appropriate for the development of robust and cost-effective markers to infer the phylogeny of genus *Nicotiana* and family Solanaceae. Our comparative plastid genome analysis revealed that the maternal parent of the tetraploid *N. rustica* was the common ancestor of *N. paniculata* and *N. knightiana*, and the later species is more closely related to *N. rustica*. The relaxed molecular clock analyses estimated that the speciation event between *N. rustica* and *N. knightiana* appeared 0.56 Ma (HPD 0.65–0.46). The biogeographical analysis showed a south-to-north range expansion and diversification for *N. rustica* and related species, where *N. undulata* and *N. paniculata* evolved in North/Central Peru, while *N. rustica* developed in Southern Peru and separated from *N. knightiana,* which adapted to the Southern coastal climatic regimes. We further inspected selective pressure on protein-coding genes among tobacco species to determine if this adaptation process affected the evolution of plastid genes. These analyses indicated that four genes involved in different plastid functions, such as DNA replication (*rpo*A) and photosynthesis (*atpB*, *ndh*D and *ndhF*), came under positive selective pressure as a result of specific environmental conditions. Genetic mutations of the following genes might have contributed to the survival and better adaptation during the evolutionary history of tobacco species.

Key words: *Nicotiana*, Chloroplast genome, Substitution and InDels, Mutational hotspots, substitutions, positive selection.

1. Introduction

The plant family Solanaceae consists of 98 genera and ~ 2700 species (Olmstead et al., 2008; Olmstead & Bohs, 2007). This megadiverse family consists of herbaceous annual species to perennial trees with a natural distribution ranging from deserts to rainforests (Knapp et al., 2004). *Nicotiana* L. is the fifth largest genus in the family, comprising 76 species, which were subdivided into three subgenera and fourteen sections by Goodspeed (1954). The subgenera of *Nicotiana*, as proposed by Goodspeed (1954), were not monophyletic (Aoki & Ito, 2000; Chase et al., 2003), but most of Goodspeed's sections were natural groups. The formal classification of the genus has been refined to reflect the growing body of evidence on *Nicotiana*, consisting of thirteen sections (Knapp, Chase & Clarkson, 2004). Most *Nicotiana* species are diploid (2n = 2x = 24) while allopolyploid species are also reported (Leitch et al., 2008). Phylogenetic studies have shown that these allopolyploids were formed 0.2 million (*N. rustica* L. and *N. tabacum* L.) to more than 10 million years ago (species of sect. *Suaveolentes*) (Clarkson et al., 2004; Leitch et al., 2008). Cultivated tobacco (*N. tabacum* L.), commonly grown for its leaves and an important economic and agricultural crop around the world (Occhialini et al., 2016), is a natural amphiploidy derived from two progenitors (Smith 1974). *Nicotiana* species, especially *N. tabacum*, are also used as model organisms in plant sciences and genetics (Zhang et al., 2011). The first complete chloroplast genome sequence was also published for this species (Shinozaki et al., 1986). Since the publication of this sequence, the structure and composition of chloroplast genomes has become widely utilized in identifying unique genetic changes and the evolutionary relationships of various groups of plants, while plastid genes have also been linked with important crop traits such as yield and resistance to various pest and pathogens (Jin & Daniell, 2015). Chloroplasts (cp) are large double membrane organelles with a genome size of 75-250 kb (Palmer, 1985). Proteins are used not only for photosynthesis but also for the synthesis of fatty acids and amino acids (Cooper, 2000). Angiosperm plastomes commonly contain ~130 genes with up to 80 protein-coding, 30 transfer RNA (tRNA), and four ribosomal RNA (rRNA) genes (Daniell et al., 2016). The plastid genome exists in circular and linear forms (Oldenburg & Bendich, 2015) and the percentage of each form varies within plant cells (Oldenburg & Bendich, 2016). Circular formed plastomes have a typical quadripartite structure, with two inverted repeat regions (IRa and IRb), separated by one large single-copy (LSC) and one small single-copy

(SSC) region (Palmer, 1985; Amiryousefi, Hyvönen & Poczai, 2018a; Abdullah et al., 2019b). Numerous mutational events occur in plastid genomes: variations in tandem repeats, insertion and deletions (indels), and point mutations, but inversions and translocations are also common (Jheng et al., 2012; Xu et al., 2015; Abdullah et al., 2019a). The plastid genome of angiosperms have a uniparental maternal inheritance (Daniell, 2007) but paternal inheritance has been recorded in a few gymnosperm species (Neale & Sederoff, 1989). The conserved organization of the plastid genome makes it extremely useful in exploring the phylogenetic relationships at various taxonomic levels (Ravi et al., 2008). Polymorphism in the chloroplast genome has been exploited to solve taxonomic issues, infer phylogeny and to investigate species adaptation to their natural habitats (Daniell et al., 2016). Genes in the plastid genome encode proteins and several types of RNA molecules, which play a vital role in functional plant metabolism, and can consequently undergo selective pressures. Most plastid protein-coding genes are under purifying selection to maintain their function, while positive selection might act on some genes in response to environmental changes. Complete plastid genome sequences are also useful tools in population genetics (Ahmad, 2014), species barcoding (Nguyen et al., 2017) transplastomic (Waheed et al., 2011, 2015) and conservation of endangered species (Wambugu et al., 2015).

Here, we assembled the plastid genome of five *Nicotiana* species and compared their sequences to gain insight into the chloroplast genome structure of the genus *Nicotiana*. We also inferred the phylogenetic relationship of genus *Nicotiana* and investigated the selection pressures acting on protein-coding genes, then identified mutational hotspots in the *Nicotiana* plastid that might be used for the development of robust and cost-effective markers in crop breeding or taxonomy.

2. Materials and Methods

2.1. Chloroplast genomes assembly and annotation

Illumina sequence data of *Nicotiana knightiana* L. (13.1 Gb, accession number SRR8169719*)*, *N. rustica (*15.5 Gb, SRR8173839*)*, *N. paniculata* (35.1 Gb, SRR8173256*)*, *N. obtusifolia (*23 Gb, SRR3592445*)* and *N. glauca (*12.5 Gb, SRR6320052*)* were downloaded from the Sequence Read Archive (SRA). The chloroplast genome sequence contigs were selected by performing the BWA alignment with default settings (Li & Durbin, 2009) using *Nicotiana tabacum* (GenBank accession number: NC_001879) as a reference. Geneious R8.1 *de novo* assembler (Kearse et al., 2012) was used to order the selected contigs for final assembly. The genome sequence was

annotated using GeSeq (Tillich et al., 2017) and CPGAVAS2 (Shi et al., 2019). Following *de novo* annotation, start/stop codons and the position of introns were manually inspected and curated. The tRNA genes were verified by tRNAscan-SE version 2.0 with default settings (Lowe & Chan, 2016) and Aragorn version 1.2.38 (Laslett & Canback, 2004). Circular genome maps were drawn with OGDRAW v1.3.1 (Greiner, Lehwark & Bock, 2019). The average coverage depth of *Nicotiana* species plastid genomes was determined by mapping all reads to *de novo* assembled plastid genomes with BWA (Li & Durbin, 2009) visualized with Tablet (Milne et al., 2009). Novel *Nicotiana* plastid genomes were deposited in NCBI and the assigned accession numbers are shown in Table 1.

2.2. Comparative genome analysis and RNA editing prediction

Novel plastid genome sequences were compared through multiple alignments using MAFFT v7 (Katoh & Standley, 2013). Every part of the genome, such as intergenic spacer regions (IGS), introns, protein-coding genes, and ribosomal RNAs and tRNAs, was considered for comparison. Each part was extracted and used to determine nucleotide diversity in DnaSP v6 (Rozas et al., 2017). Substitution, transition and transversion rates were also calculated compared to the *N. tabacum* reference using Geneious R8.1 (Kearse et al., 2012). Structural units of the plastid genome (LSC, SSC and IR) were individually aligned to determine the rate of substitutions and to further search for indels using DnaSP v6. The expansion and contraction of inverted repeats and their border positions were compared for ten selected *Nicotiana* species using IRscope (Amiryousefi, Hyvönen and Poczai, 2018b). The online software PREP-cp (Putative RNA Editing Predictor of Chloroplast) was used with default settings to determine putative RNA editing sites (Mower, 2009) and the codon usage and amino acids frequencies were determined by Geneious R8.1 software (Kearse et al., 2012).

2.3. Repeats analyses

Microsatellites repeats within the plastid genomes of five *Nicotiana* species were detected using MISA (Beier et al., 2017) with the minimal repeat number of 7 for mononucleotide repeats, 4 for di- and 3 for tri-, tetra-, penta- and hexanucleotide SSRs. We also used REPuter software (Kurtz, 2002) with the following parameters: minimal repeats size was set to 30 bp, Hamming distance to 3, minimum similarity percentage of two repeats copies up to 90%, maximum computed repeats numbers to 500 bp for scanning and visualizing forward (F), reverse (R), palindromic (P)

and complementary (C) repeats. Tandem repeats were found with the tandem repeats finder using default parameters (Benson, 1999).

2.4. Synonymous (K_s) and non-synonymous (K_a) substitution rate analysis

The synonymous (K_s) and non-synonymous (K_a) substitution were analyzed using the chloroplast genome of *Nicotiana tabacum* as reference for all *de novo* assembled *Nicotiana* plastid genomes. For this purpose, protein-coding genes were extracted from *Nicotiana* plastomes, then aligned with the corresponding genes of *S. dulcamara* as a reference using MAFFT (Katoh & Standley, 2013) and analyzed using DnaSP software (Rozas et al., 2017). We further assessed the impact of positive selection using additional codon models to estimate the rates of synonymous and nonsynonymous substitution. The signs of positive selection were further assessed using fast unconstrained Bayesian approximation (FUBAR) (Murrell et al., 2013) and the mixed effects model of evolution (MEME) (Murrell et al., 2012) as implemented in the DATAMONKEY web server (Delport et al., 2010). Sites with cut-off values of PP < 0.9 in FUBAR were considered as candidates to have evolved under positive selection. Out of all analyses performed in DATAMONKEY, the most suited model of evolution for each data set, directly estimated on this web server, was used. In addition, the mixed effects model of evolution (MEME), a branch-site method incorporated in the DATAMONKEY server, was used to test for both pervasive and episodic diversifying selection. MEME applies models variable ω across lineages at individual sites, restricting ω to be ≤ 1 in a proportion p of branches and unrestricted at a proportion (1 − p) of branches per site. Positive selection was inferred with this method for a P value < 0.05.

2.5. Phylogenomic analyses

Plastid genome sequences from the genus *Nicotiana* were selected from Organelle Genome Resources of NCBI, accessed on 21.2.2019. We included all available plastid genome sequences of tobacco species in our analysis and added *de novo* assembled sequences while *S. dulcamara* was used as an outgroup. For the species included in our analysis, coding alignments were constructed from the excised plastid genes using MACSE (Ranwez et al., 2011), including the following seventy-five protein coding genes: *atp*A, B, E, F, H, I; *ccs*A; *cem*A; *clp*P; *mat*K; *ndh*A, B, C, D, E, F, G, H, I, J, K; *pet*A, B, D, G, J, L, N; *psa*A, C, I, J; *psb*A, B, C, D, E, F, I, L, M, N, T, Z; *rbc*L; *rpl*2, 7, 14, 16, 19, 20, 22, 23, 32, 33, 36; *rpo*A, B, C1; *rps*2, 3, 4, 7, 8, 11, 14, 15, 16, 18, 19; *ycf*2, 3, 4. For phylogenetic analysis we used a matrix of protein-coding genes of

twelve species with a concatenated matrix length of 75,449 bp. The best fitting model (GY+F+I+G4) was determined by ModelFinder (Kalyaanamoorthy et al., 2017) as implemented in IQ-TREE according to the Akaike information criterion (AIC), and Bayesian information criterion (BIC). Maximum likelihood (ML) analyses were performed with IQ-TREE (Nguyen et al., 2015) using the ultrafast bootstrap approximation (UFBoot; Hoang et al., 2018) with 1,000 replicates and the SH-like approximate likelihood ratio test (SH-aLRT), also with 1,000 bootstrap replicates, and TreeDyn was used for further enhancement of phylogenetic tree analysis (Dereeper et al., 2008; Lemoine et al., 2019).

Relative divergence times were estimated for the species *N. rustica* and putative parental species using BEAST v.1.8.4 (Drummond et al., 2012), applying GTR + I + G rate substitution to the protein-coding plastid gene matrix. A Yule speciation tree prior and a relaxed uncorrelated clock- model that allows rates to vary independently along branches (Drummond et al., 2006) were used, with all other parameters set to default. The median time split between the *S. dulcamara* and *N. undulata* (mean = 25 Myr; standard deviation = 0.5) was used as a temporal constraint to calibrate the BEAST analyses derived from the Solanaceae-wide phylogeny of Särkinen et al. (2013) and the Time Tree of Life (Kumar et al. 2017). Uncertainty regarding these dates was incorporated by assigning normal prior distributions to the two calibration points (Couvreur et al., 2008; Evans et al., 2014). Four independent BEAST runs were conducted, each with 10 million generations, sampling every 10,000 generations. Convergence of all parameters was assessed in Tracer 1.5 (Rambaut et al., 2014) and 10% of each chain was removed as burn-in. The Markov chains were combined in LogCombiner 1.7.2. (Drummond et al., 2012) to calculate the maximum clade credibility tree.

We defined six biogeographical areas based on Köppen-Geiger climatic and further biogeographic evidence and distributions: (A) Colombian/Ecuadorian mountain range mixed equatorial (Af), monsoon (Am) and temperate oceanic climate (Cfb), (B) Northern Peruvian mountain range with tropical savanna climate (Aw), (*C*) Central Peru with equatorial climate (*Af*), (*D*) Coastal Peru with cold semi-arid and desert climate (*Bsk*, *BWk*), (*E*) Peruvian Mountain range with humid subtropical/oceanic highland climate (*Cwb*), (*F*) Bolivian/Chilean alpine/mountain range with mixed semi-arid cold (*Bsk*, *BWk*) and humid subtropical climate (*Cwa*). These areas were used in the Bayesian Binary Method (BBM) model implemented in RASP (Yu, Blair & He, 2019) to investigate the biogeographic history of the selected four

Nicotiana species. BBM infers ancestral area using a full hierarchical Bayesian approach and hypothesizes a special "null distribution", meaning that an ancestral range contains none of the unit areas (Ronquist 2004). The analysis was performed on the BEAST maximum clade credibility tree using default settings, i.e. fixed JC + G (Jukes-Cantor + Gamma) with null root distribution. Ancestral area reconstruction for each node was manually plotted on the BEAST tree using pie charts. Species distributions were determined from data stored in the Solanaceae Source Database (http://solanaceaesource.org/) and Global Biodiversity Information Facility (GBIF) (https://www.gbif.org/).

3. RESULTS

3.1. Characteristics of *Nicotiana* plastid genomes

Five *Nicotiana* species chloroplast genomes were assembled and the lengths of these plastid genomes were: *Nicotiana knightiana* (155,968 bp), *Nicotiana rustica* (155,849 bp), *Nicotiana paniculata* (155,689 bp), *Nicotiana obtusifolia* (156,022 bp) and *Nicotiana glauca* (155,917 bp). Further details of the characteristics of the assembled plastid genomes are summarized in Table S1. The coverage of assembled plastid genomes was 811× *Nicotiana knightiana*, 1,951× *Nicotiana rustica*, 1,032× *Nicotiana paniculata*, 1,412× *Nicotiana obtusifolia* and 327× *Nicotiana glauca*. The GC content of IR regions were highest (43.2%) followed by LSC (35.9%) and SSC (32.1%) (Table S1). The high GC content of IR was due to high GC content of the tRNAs (52.9%) and rRNAs (55.4%) genes.

De novo assembled *Nicotiana* plastid genomes had 134 unique genes, whereas eighteen genes were duplicated in the IR region (Table S2, Fig.1). Out of 134 genes, 86 were protein-coding genes, 37 were tRNA genes and 8 were rRNA genes. Among 18 duplicated genes in IR region, 7 were protein-coding, 7 were tRNA genes and 4 were rRNA genes. 18 intron-containing genes were present in the plastome of *Nicotiana* species. The *rps12* gene is a trans-spliced gene, its 1st exon existing in the LSC region while the 2nd and 3rd exons are in the IR region.

3.2. Comparative analyses, codon usage and RNA editing sites

The nucleotide composition of *Nicotiana* species was compared, and all genomes had similar nucleotide composition indicating high synteny in the LSC, SSC, IR and CDSs but also in non-coding regions. Detailed comparison of the base composition is shown in Table S3. A high percentage of hydrophobic amino acids were encoded in *Nicotiana* plastid genomes, while acidic

amino acids were present at lower rates. The amino acids are AT rich sequences as compared to GC (Fig. 2A). Relative synonyms codon usage (RSCU) and frequency of amino acid revealed that leucine is the most abundant and cysteine was the least encoded amino acid in these genomes (Fig S1). The codon usage revealed a high frequency of codons with A/T at 3rd codon position as compared to C/G at 3rd codon position (Table S4).

The number of predicted RNA editing sites using PREP-cp varied between 34 and 37, distributed among fifteen genes (see Table S3). Among these genes, *ndhB* (9) possessed the most of these sites, followed by *ndhD* (6-8) and *rpoB* (4). The *ndhD* gene revealed a fraction of variation among species: *N. knightiana, N. rustica* and *N. paniculata* having six RNA editing sites whereas seven were observed in *N. obtusifolia* and eight in *N. glauca*. Most of the RNA editing sites were C to U edits on the first and second base of the codons, but the frequency of second base codon edits was much higher. The conversions from serine to leucine were the most frequent and these changes helped in the formation of hydrophobic amino acids, i.e. valine, leucine and phenylalanine (Table S5).

3.3. IR contraction and expansion

The LSC/IR and IR/SSC border positions of *Nicotiana* plastid genomes were compared (Fig 3) using IRscope. The length of the IR regions was similar, ranging from 25,331bp to 25,436bp showing some expansion. The endpoint of the Solanaceae JLA (IRa/SSC) is characteristically located upstream of the *rps*19 and downstream of the *trn*H-GUG, which was confirmed in *Nicotiana*. In *N. tomentosiformis*, the IR expanded to partially include *rps*19, creating a truncated ψ*rps*19 copy at JLA, which was thought to be missing from the entire *Nicotiana* clade (Amiryousefi, Hyvönen & Poczai, 2018a). In this species the IR region has expanded to include 60 bp of *rps*19. The extent of the IR expansion to *rps*19 varied from 2 to 60 bp and the end point seems to be conserved to the following intergenic spacer region. Furthermore, *inf*A, *ycf*15, and a copy of *ycf*1 located on the JSB were detected as pseudogenes. The position of *ycf1* in the IRb/SSC region varied. It left a 36 bp pseudogene in *N. knightiana, N. rustica* and *N. glauca*, 33 bp pseudogene in *N. obtusifolia* and a 72 bp one in *N. paniculata.*

3.4. Non-synonymous (K_a) and synonymous (K_s) substitution rate analysis

Synonymous/non-synonymous substitutions ratio is widely used as an indicator of adaptive evolution or positive selection (Kimura, 1979). We have calculated the K_s, K_a and K_a/K_s ratio for

77 protein-coding genes for five selected *Nicotiana* species using *S. dulcamara* as a reference (Table S6). Among the analyzed genes, 31 had K_s=0, 19 had K_a=0, and 39 genes had both K_s and K_a=0 values. Of the investigated genes, 21 showed a K_a/K_s ratio of more than 0.5. Eight of these genes (*atpF, psaA, ycf4, psbB, infA, ndhB, rpl32* and *ccsA)* had a K_a/K_s ratio greater than 0.5 for one species, *ycf1* had K_a/K_s greater than 0.5 for two species, while *atpA, rps2, rpoB, rps12, ycf2, ndhG* had K_a/K_s greater than 0.5 for three species whereas genes *rpoC1, atpB, rpoA, ndhD* had K_a/K_s ratio more than 4 species and *rpoC2* and *ndhF* had K_a/K_s ratio for all species. We selected the genes *atp*B, *rpo*A, *ndh*D, *ndh*F, *rpo*C1 and C2 for further analysis using FUBAR and MEME. FUBAR estimates the number of nonsynonymous and synonymous substitutions at each codon given a phylogeny, and provides the posterior probability of every codon belonging to a set of classes of ω (including $\omega = 1$, $\omega < 1$ or $\omega > 1$) (Murrell et al., 2013). MEME estimates the probability for a codon to have undergone episodes of positive evolution, allowing the ω ratio distribution to vary across codons and branches in the phylogeny. This last attribute allows identification of the proportion of codons that may have been evolving neutrally or under purifying selection, while the remaining codons can also evolve under positive selection (Murrell et al., 2012). The two models indicated positive selection on the codons only found in *atp*B, *rpo*A, *ndh*F and *rpo*A (Table 1). Thus, the methods described suggested six amino acid replacements altogether as candidates for positive selection, of which three were fixed in all *Nicotiana*, and three were restricted to diverse groups of species (see Table 1).

3.5. Repetitive sequences in novel *Nicotiana* plastid genomes

Repeat analysis performed with MISA revealed high similarity in chloroplast microsatellites (cpSSRs) ranging from 368 to 384 among tobacco species. The majority of the SSRs in these plastid genomes were mononucleotide rather than trinucleotide or dinucleotide. The most dominant of the SSRs were A/T motifs mononucleotides, and in dinucleotides AT/TA motifs were the second most predominant. Mononucleotide SSRs varied from 7-17 units repeats; dinucleotide SSRs from 4-5-unit repeats while other SSRs types were present mainly in 3-unit repeats. Mostly the SSRs existed in LSC, in comparison to IR and SSC (Fig 4) (Table S7). REPuter software was used to identify and locate forward (F), reverse (R), palindromic (P), and complementary (C) repeats in all the species of *Nicotiana*. In the plastomes of five *Nicotiana* species, we found 117 oligonucleotide repeats: 25 in *N. knightiana,* 23 in *N. rustica,* 21 in *N. paniculata,* 23 in *N. obtusifolia,* 25 in *N. glauca.* Forward (F) and palindromic repeats were

present in large numbers as compared to others in all species: 11 (44%) (F) and 14 (56%) (P) in *N. knightiana*, 14 (60%) (F) and 9 (39%) (P) in *N. rustica*, and 12 (57%) (F) and 9 (42%) (P) in *N. paniculata*, 14 (56%) (F) and 11 (44%) (P) in *N. obtusifolia*, 9 (39%) (F) and 11 (52%) (P) in *N. glauca*. The size of oligonucleotide repeats varied from 30-65 bp, and many of these repeats were 30-35 bp in length. The LSC region held most of the identified oligonucleotide repeats as compared to SSC and IR. The LSC region contained 13 in *N. knightiana*, 11 in *N. rustica*, 14 in *N. paniculata*, 15 in *N. obtusifolia* and 17 in *N. glauca*. In plastid genome regions, the repeats existed mostly in IGS, followed by CDS and intronic regions (Fig. 5) (Table S8). The number of tandem repeats varies from 24-27 between these *Nicotiana* species. The IGS region contains the most tandem repeats followed by the CDS region. The size of these repeats varied between 20 to 88 among *Nicotiana* (Fig. 6).

3.6. Single nucleotide polymorphism and insertion/deletion analyses in *Nicotiana*

We investigated substitution types in the five plastomes of *Nicotiana* species (one IR removed), using *Nicotiana tabacum* as a reference. *Nicotiana knightiana* (786), *Nicotiana rustica* (775), *Nicotiana paniculata* (861), *Nicotiana obtusifolia* (847) and *Nicotiana glauca* (509) substitutions were seen in the whole plastid genome. The types of substitutions exhibited among *Nicotiana* species were similar. Most of the conversions were A/G and C/T in comparison to other SNPs (single nucleotide polymorphism) (Table 2). Ts/Tv ratio were as follows: *Nicotiana knightiana* LSC (1.5), SSC (0.968) and IR (1.047), *Nicotiana rustica* LSC (1.496), SSC (0.978) and IR (1), *Nicotiana paniculata* LSC (1.461), SSC (0.886) and IR (0.833), *Nicotiana obtusifolia* LSC (1.097), SSC (1.020) and IR (1.194), *Nicotiana glauca* LSC (0.924), SSC (0.819) and IR (0.783) (Table S9). The substitutions in different regions of these genomes are *Nicotiana knightiana* contains 560 (LSC), 43 (IR) and 183 (SSC) SNPs, *Nicotiana rustica* contains 599 (LSC), 32 (IR) and 183 (SSC) substitutions, *Nicotiana paniculata* has 630 (LSC), 33 (IR) and 198 (SSC) substitutions, *Nicotiana obtusifolia* consists of 671 (LSC), 68 (IR) and 210 (SSC) substitutions while *Nicotiana glauca* has 327 (LSC), 82 (IR) and 100 (SSC). Insertions and Deletions (indels) were also examined using DnaSP in all regions of the chloroplast genome. In total, *Nicotiana knightiana* (110), *Nicotiana rustica* (107), *Nicotiana paniculata* (116), *Nicotiana obtusifolia* (143) and *Nicotiana glauca* (113) indels were found. The LSC region held the majority of the indels, followed by SSC, whereas IR contained minimum indels (Table 3).

3.7. Divergence hotspot regions in *Nicotiana*

The CDS, intron and IGS regions of the whole plastid genome of five *Nicotiana* species were compared to discover polymorphic regions (mutational hotspots). High polymorphism was found in intronic regions (average π=0.167) in comparison to IGS (π=0.031) and CDS regions (average π=0.002). Among *Nicotiana* species, the nucleotide diversity values varied from 0 (*ycf3*) to 0.306 (*rps12 intron* region) (Fig. 7). Here, 20 highly polymorphic regions were determined that might be used as potential makers to reconstruct the phylogeny for identifying *Nicotiana* species (Table 4).

3.8. Phylogenomic analyses

Phylogenetic analysis within *Nicotiana* plastid genomes were reconstructed with the maximum likelihood method, based on selected and concatenated protein-coding genes. Our phylogenetic analyses resulted in a highly resolved tree (Fig 8), with almost all clades recovered having maximum branch support values. After the elimination of indels, the tree was reconstructed based on alignment size of 75,449 bp with the best fitting model GY+F+I+G4 (Fig 8). We further concentrated on the species phylogeny of *N. rustica* and putative parental species where relative divergence times were estimated using a relaxed uncorrelated clock implemented in BEAST. This analysis found that the divergence of *N. undulata* appeared 5.36 (highest posterior density, HPD 6.38–4.43) million years ago (Ma), while *N. paniculata* diverged 1.17 Ma (HPD 2.18–0.63) followed by the most recent split of *N. rustica* and *N. knightiana* 0.56 Ma (HPD 0.65–0.46). This analysis showed that the *Nicotiana* species included in the analysis are not older than the end of the Pliocene and that most subsequent evolution must have occurred in the Pleistocene. The timing of these lineage splits, in addition to the current distributions of four closely related species, were used to infer the progression of migratory steps in RASP (Fig 9). The most recent common ancestor (MRCA) area illustrated a dispersal event for *N. paniculata* in Northern (B) and Southern Peru (E) and the vicariance of *N. knightiana* in Coastal Peru (D). The overall dispersal pattern of the examined species showed a south-to-north expansion pattern from Central Peru to Colombia and Ecuador (*N. rustica*) to Bolivia (*N. undulata*).

4. Discussion

4.1. Molecular evolution of *Nicotiana* plastid genomes

We compared five chloroplast genomes of *Nicotiana* species, which revealed similar genomic features. These comparative analyses produced an insight into the phylogeny and evolution of *Nicotiana* species. The GC content of the *Nicotiana* species referred to above were similar to those of other *Nicotiana* species (Sugiyama et al., 2005; Yukawa, Tsudzuki & Sugiura, 2006) i.e. the GC content in the IR is high, which might be a result of the existence of ribosomal RNA (Qian et al., 2013; Cheng et al., 2017; Zhao et al., 2018). The genome organization, gene order and content and of these *Nicotiana* species were also similar for *N. slyvestris* and *N. tabacum* (Sugiyama et al., 2005; Yukawa, Tsudzuki & Sugiura, 2006). The intron plays an important role in the regulation of gene expression (Xu et al., 2003). The *trn*K intron is important because it expresses an unusual form of a group II intron derived from a mobile group of mitochondrial-like intron open reading frames (ORFs) (Hausner et al., 2006). As in the plastid genomes of many land plants, the abundance of A/T content at 3rd base of codons was reported due to high concentration of A/T nucleotides in the whole plastid genome (Menezes et al., 2018).

The plastomes of land plants have conserved structure but diversity prevails at the border position of LSC/SSC/IR of the genome. The size range of LSC, SSC and IR varies between the plastid genomes of species that advances to alterations in several genes and leads to the deletion of one copy of a gene or duplication of functional and non-functional genes of different sizes (Menezes et al., 2018; Saina et al., 2018). In the current study, all these ten *Nicotiana* species showed similarities with some variation as compared to the *Nicotiana tabacum:* in all these plants except *Nicotiana tomentosiformis* which have 60 bp in IRb region, the *rps19* gene is present entirely in the LSC region but in *Nicotiana tabacum rps19* gene extended 5bp in the IR region. The fluctuations at the border positions of various regions of the plastid genome might be helpful in determining the evolution of species (Menezes et al., 2018). Liu et al., (2018) reported that the similarities at the junction regions may be useful in explaining the relationship between the species and that those plants which have a high level of relatedness show minimal fluctuations at the junctions of the chloroplast genome. The resemblance at junctions reveals a close relationship between the *Nicotiana* species.

Repeats in the chloroplast genome are useful in evolutionary studies and play a vital role in genome arrangement (Zhang et al., 2016). Here, we detected that the mononucleotide repeats (A/T), and trinucleotide SSRs (ATT/TAA) were present in large amounts in all the species of *Nicotiana*, which may be a result of the A/T rich proportion of chloroplast genome. A similar

result was also reported in *Nicotiana otophora* (Asaf et al., 2016). In all the species of *Nicotiana*, the LSC region contained a greater amount of SSRs in comparison to SSC and IR, which has also been demonstrated in other studies of angiosperm plastomes (Shahzadi et al., 2019; Mehmood et al., 2019). When less genomic resources are available for revealing divergence hotspot regions, the oligonucleotide repeats might be utilized as a substitute for identifying polymorphic regions (Ahmed et al., 2012; Ahmad, 2014). The current study results of oligonucleotide repeats are similar to the previously reported results of *Nicotiana* species and other angiosperm plastome studies (Asaf et al., 2016; Yang et al., 2019). Thus, the presence of both the high divergence regions in IGS and oligonucleotide repeats suggest that these regions are suitable for the development of markers to demonstrate phylogenetics relationships.

To understand the molecular evolution, it is important to know about the nucleotide substitution rates (Muse & Gaut, 1994). LSC and SSC regions are more prone to substitutions and indels whereas the IR regions are more conserved in the chloroplast genome (Ahmed et al., 2012; Abdullah et al., 2019b). Our results also showed similar results in that the IR region is mostly conserved, and most of the substitutions occurs in the LSC and SSC regions. Thus, the ratio (Ts/Tv) was equal to or more than 1. Similar results were shown in the chloroplast genome of *Dioscorea polystachya* (Yam) (Cao et al., 2018).

Divergence hotspot regions of the plastid genome could be used to develop accurate, robust and cost-effective molecular markers for population genetics, species barcoding and evolutionary based studies. (Ahmed et al., 2013; Ahmad, 2014; Nguyen et al., 2017). Previously, in several studies, polymorphic loci were identified based on comparisons of chloroplast genome to provide information about suitable loci for the development of molecular markers (Choi, Chung & Park, 2016; Li et al., 2018; Menezes et al., 2018). We found 20 polymorphic regions such as *infA, rps12 intron, rps16-trnQ-UUG* which have 0.25942, 0.15275, 0.08451 nucleotide diversity respectively that were more polymorphic than frequently used markers such as *rbcL*, and *matK*. These regions could be suitable markers for population genetics and phylogenetic analyses, especially in the genus *Nicotiana*.

4.2. Positive selection on *Nicotiana* plastid genes

Plants have evolved complex physiological and biochemical adaptations to adjust and adapt to a variety of environmental stresses. *Nicotiana*, originating in South America, has spread to many

regions of the world and members of the genus have successfully adapted to harsh environmental conditions to survive. This great variation in their distributional range induced distinctive habits and morphology in the inflorescence and flowers, indicative of the physiological specialization to the area where they evolved. Desert ephemeral *Nicotiana* species are short while subtropical perennials have tall and robust habits with variable inflorescences ranging from pleiochasial cymes to solitary flowers and diffuse panculate-cymose mixtures. For example, members of *Nicotiana* section *Suaveolentes* Goodsp. evolving in isolation faced several cycles of harsh climate change. In Australia, the native range of the species, a predominantly warm and wet environment went through intensive aridification (Poczai, Hyvönen & Symon, 2011). Throughout this climate change and increasing central aridification, many species either retreated to the wetter coastline or adapted to and still survive in this hostile inland environment (Bally et al., 2018). Tobacco plants also developed specialized biosynthetic pathways and metabolites, such as nicotine, which serve complex functions for ecological adaptations to biotic and abiotic stresses, most importantly serving as a defense mechanism against herbivores (Xu et al. 2017). Therefore, *Nicotiana* is a rich reservoir of genetic resources for evolutionary biological research, since several members of the genus went through changing climatic events and adopted to environmental fluctuations.

The patterns of synonymous (K_s) and non-synonymous (K_a) substitution of nucleotides are essential markers in evolutionary genetics defining slow and fast evolving genes (Kimura, 1979). K_a/K_s values >1, =1, and <1 indicate positive selection, natural evolution and purifying selection, respectively (Lawrie et al., 2013), while a minimal ratio of K_a/K_s (<0.5) in many genes represents purifying selection working on them. Many proteins and RNA molecules encoded by the chloroplast genomes are under purifying selection since they are involved in important functions of plant metabolism, self-replication and photosynthesis and therefore play a pivotal role in plant survival (Piot et al., 2018). Departure from the main purifying selection in case of plastid genes might happen in response to certain environmental changes when advantageous genetic mutations might contribute to survival and better adaptation. The K_a/K_s ratios in our analysis for *Nicotiana* species indicated changes in selective pressures. The genes *atp*B, *ndh*D, *ndh*F, *rpo*A, *rps*2 and *rps*12 had greater K_a/K_s value (> 1), possibly due to positive selective pressure as a result of specific environmental conditions. This has been conclusively supported by an integrative analysis using Fast Unconstrained Bayesian AppRoximation (FUBAR) and

Mixed Effects Model of Evolution (MEME) methods, which identified the set of positively selected codons in case of *atp*B, *ndh*D, *ndh*F and *rpo*A (Table 5), but provided no further evidence for *rps*2 and *rps*12.

These genes are involved in different plastid functions, such as DNA replication (*rpo*A) and photosynthesis (*atpB*, *ndh*D and *ndhF*). The *rpo*A gene encodes the alpha subunit of PEP, which is believed to predominantly transcribe photosynthesis genes (Hajdukiewicz, Allison & Maliga, 1997). The transcripts of plastid genes encoding the PEP core subunits are transiently accumulated during leaf development (Kusumi et al., 2011), thus the entire *rpo*A polycistron is essential for chloroplast gene expression and plant development (Zhang et al., 2018). The housekeeping gene *atp*B encodes the *β*-subunit of the ATP synthase complex, which has a highly conserved structure that couples proton translocation across membranes with the synthesis of ATP (Gatenby, Rothstein & Nomura, 1989), which is the main source of energy for the functioning of plant cells. In chloroplasts, linear electron transport mediated by PSII and PSI produces both ATP and NADPH, whereas PSI cyclic electron transport preferentially contributes to ATP synthesis without the accumulation of NADPH (Peng & Shikanai, 2011). Chloroplast NDH monomers are sensitive to high light stress, suggesting that the *ndh* genes encoding the NAD(P)H dehydrogenase (NDH) may also be involved in stress acclimation through the optimization of photosynthesis (Casano, Martín & Sabater, 2001; Martin et al., 2002; Rumeau, Peltier & Cournac, 2007). During acclimation to growth light environments, many plants change biochemical composition and morphology (Terashima et al., 2005). The highly responsive regulatory system controlled by cyclic electron transport around PSI could optimize photosynthesis and plant growth under naturally fluctuating light (Yamori, 2016). When the demand for ATP is higher than that for NADPH (e.g., during photosynthetic induction, at high or low temperature, at low CO_2 concentration, or under drought), cyclic electron transport around PSI is likely to be activated (Yamori, 2016; Yamori & Shikanai, 2016). Thus, positive selection acting on ATP synthase and NAD(P)H dehydrogenase encoding genes is probably evidence for adaptation to novel ecological conditions in *Nicotiana*.

These findings might be also supported by our observation that RNA editing sites occurred frequently in *Nicotiana ndh* genes (Table S3). It has been shown that *ndh*B mutants under lower air humidity conditions or following exposure to ABA present a reduction in the photosynthetic level, likely mediated through stomatal closure triggered under these conditions (Horvath et al.,

2000). Therefore, a protein structure modification resulting from a loss or decrease in RNA editing events could affect adaptations to stress conditions or cause other unknown changes (Rodrigues et al., 2017). Previous studies have demonstrated that abiotic stress influences the editing process and consequently plastid physiology (Nakajima & Mulligan, 2001). Alterations in editing site patterns resulting from abiotic stress could be associated with susceptibility to photo-oxidative damage (Rodrigues et al., 2017) and indicate that *Nicotiana* species experienced abiotic stresses during their evolution, which resulted in positive selection of some of the plastid genes. Up to this point, positive selection has rarely been detected in chloroplast genes except for *clp*P1 (Erixon & Oxelman, 2008), *ndh*F (Peng et al., 2011), *mat*K (Hao, Chen & Xiao, 2010) and *rbc*L (Kapralov et al., 2011). However, a recent study by Piot et al. (2018) showed that one-third of the plastid genes in 113 species of grasses (Poaceae) evolved under positive selection. This might indicate that positive selection might be overlooked among diverse groups of plant taxa.

4.3. Phylogenetic relationships and the origin of tetraploid *Nicotiana rustica*

Our comparative plastid genome analysis revealed that the maternal parent of the tetraploid *N. rustica* was the common ancestor of *N. paniculata* and *N. knightiana*, and the later species is more closely related to *N. rustica*. The relaxed molecular clock analyses estimated that the speciation event between *N. rustica* and *N. knightiana* appeared ~0.56 Ma (HPD 0.65–0.46) in line with previous findings (Sierro et al., 2018). Comparative analysis of the genomes of four related *Nicotiana* species revealed that *N. rustica* inherited about 41% of its nuclear genome from its paternal progenitor, *N. undulata*, the rest from its maternal progenitor, the common ancestor of *N. paniculata* and *N. knightiana* (Sierro et al., 2018), which has also been confirmed by our study. It has been shown that *N. rustica* and in fact all *Nicotiana* tetraploids, except species included in section *Suaveolentes*, originated from a doubling of the diploid chromosome for the genus. Thus, they should be regarded as natural allopolyploids (Leitch et al., 2008). We also revealed that *N. knightiana* is more closely related to *N. rustica* than *N. paniculata*, which can be further corroborated by the distribution of indels highlighted in the present study. The biogeographical analysis carried out suggests that *N. undulata* and *N. paniculata* evolved in North/Central Peru, while *N. rustica* developed in Southern Peru and separated from *N. knightiana*, which adapted to the Southern coastal climatic regimes. Positively selected plastid genes with functions such as DNA replication (*rpo*A) and photosynthesis (*atpB*, *ndh*D and *ndhF*) might have been associated with successful adaptation to, for example, a coastal environment.

514 However, our results should be regarded as tentative, as our survey excludes several broad
515 ecological variables from testing, including variation in salinity, island versus mainland, and East
516 versus West of the Andes. We aim to highlight that many potential environmental variables
517 might be highly correlated with speciation processes, as has been demonstrated in the same
518 region for another Solanaceae group in the tomato clade (*Solanum* sect. *Lycopersicon*), where
519 amino acid differences in genes associated with seasonal climate variation and intensity of
520 photosynthetically active radiation were correlated with speciation processes (Pease et al., 2016).
521 Another example of rapid adaptive radiation from the family is the genus *Nolana* L.f., where
522 several clades gained competitive advantages in water-dependent environments by succeeding
523 and diverging in Peru and Northern Chile (Dillon et al., 2009). In the case of *N. rustica* and
524 related species we assume that diversification was driven by the ecologically variable
525 environments of the Andes. Our molecular clock analysis provides recent species diversification
526 in the Pleistocene and Pliocene while substantial climatic transitions in Peru predate these events.
527 For example, the uplift of the central region of the Andes and the formation of the Peruvian
528 coastal desert ended (~14 – 150 Mya; Hoorn et al., 2010; Gerreaud et al., 2010) before the
529 geographical and ecological expansion of *N. rustica* and related parental species.

530 The dispersal of *N. rustica* and related species shows a south-to-north range expansion and
531 diversification which has been suggested by phylogenies of other plant and animal groups in the
532 Central Andes (Picard, Sempere & Plantard, 2008; Lueber and Weigend, 2014). Based on the
533 south-to-north progression scenario, habitats located at high altitudes were first available for
534 colonization in the south, recently continuing to northward. Erosion and orogenic progression
535 caused dispersal barriers of species colonizing these high habitats to diversify in a south-to-north
536 pattern, frequently following allopatric speciation. Thus, for taxonomic groups currently residing
537 throughout a large portion of the high Andes, a south-to-north speciation pattern is expected
538 (Doan, 2003). In this case the most basal species (*N. undulata*) has more southern geographic
539 ranges, and the most derived species (*N. rustica*) has more northern geographic ranges except for
540 *N. knightiana*, which presumably colonized the coastal range of Peru. Although the four
541 *Nicotiana* species examined show overlaps in their distribution, it is probable that speciation was
542 caused by fragmentation of populations during the glacial period (see Simpson, 1975). Utilizing
543 fewer chloroplast loci for phylogenetic analyses of plant species may limit the solution of
544 phylogenetic relationships, specifically at low taxonomic levels (Hilu & Alice, 2001; Majure et

al., 2012). Previously, genus *Nicotiana* was subdivided into 13 sections using multiple chloroplast markers, i.e. *trnL* intron and *trnL-F* spacer, *trnS-G* spacer and two genes, *ndhF and matK* (Clarkson et al., 2004). Recently, inference of phylogeny based on complete chloroplast genomes has provided deep insight into the phylogeny of certain families and genera (Henriquez et al., 2014; Amiryousefi, Hyvönen & Poczai, 2018a; Abdullah et al., 2019a). Here, we reconstructed a phylogenetic tree among eleven species of genus *Nicotiana* that belong to nine sections (Clarkson et al., 2004) based on 75 protein-coding genes by using *S. dulcamara* as an outgroup which attests the previous classification of genus *Nicotiana* with high bootstrapping values. Species of each section are well resolved whereas the *N. tabacum* of section *Nicotiana* and *N. sylvestris* of section Sylvestres show close resemblance. The *N. paniculata* and *N. knightiana* belong to section Paniculatae but here did not appear on the same node. This revealed that further data is required to elucidate the phylogenetic relationship among these two species. Overall, our phylogenetic analyses support the previous classification of genus *Nicotiana,* but enrichment of chloroplast genomic resources can provide further insight into the phylogeny of the genus *Nicotiana.*

5. Conclusion

In the present study, we assembled, annotated and analyzed the whole cp genome sequence of five *Nicotiana* species. The genomic structure and organization of their chloroplast genome was like those of previously reported Solanaceae plastomes. Divergences of LSC, SSC and IR region sequences were identified, as well as the distribution and location of repeat sequences. The identified mutational hotspots sequences could be utilized as potential molecular markers to investigate phylogenetic relationships in the genus. As we demonstrated in our study to elucidate the maternal genome origins of *N. rustica*, our results could provide further help in understanding the evolutionary history of tobaccos.

Acknowledgements

We thank Kenneth Quek for editing the manuscript.

Competing interest

The authors declare that they have no conflict of interest.

Authors contributions

574 Furrukh Mehmood: Conceptualization, Genome assembly and annotation, Data analysis, Data
575 interpretation, prepared figures and tables, Manuscript drafting and editing.

576 Abdullah: Genome annotation, Data analysis, Data interpretation, Manuscript drafting.

577 Zartasha Ubaid: Data analysis, Data interpretation, Manuscript drafting.

578 Iram Shehzadi: Data analysis, Data interpretation, Manuscript drafting.

579 Ibrar Ahmed: Conceptualization, Manuscript editing.

580 Mohammad Tahir Waheed: Conceptualization, Manuscript editing.

581 Péter Poczai: Supervision, carried out selection tests and phylogenetic analysis, prepared figures
582 and tables, authored and reviewed drafts of the paper, approved the final draft.

583 Bushra Mirza: Supervision, authored or reviewed drafts of the paper, approved the final draft.

584 **References**

585 **Abdullah, Mehmood F, Shahzadi I, Waseem S, Mirza B, Ahmed I, Waheed MT. 2019a.**
586 Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): Comparative analyses and
587 identification of mutational hotspots. *Genomics* DOI: 10.1016/j.ygeno.2019.04.010.

588 **Abdullah, Shahzadi I, Mehmood F, Ali Z, Malik MS, Waseem S, Mirza B, Ahmed I,**
589 **Waheed MT. 2019b.** Comparative analyses of chloroplast genomes among three *Firmiana*
590 species: Identification of mutational hotspots and phylogenetic relationship with other
591 species of Malvaceae. *Plant Gene*:100199. DOI: 10.1016/j.plgene.2019.100199.

592 **Ahmad I. 2014.** Evolutionary dynamics in taro. Massey University, Palmerston North, New
593 Zealand.

594 **Ahmed I, Biggs PJ, Matthews PJ, Collins LJ, Hendy MD, Lockhart PJ. 2012.** Mutational
595 dynamics of aroid chloroplast genomes. *Genome Biology and Evolution* **4**:1316–1323. DOI:
596 10.1093/gbe/evs110.

597 **Ahmed I, Matthews PJ, Biggs PJ, Naeem M, Mclenachan PA, Lockhart PJ. 2013.**
598 Identification of chloroplast genome loci suitable for high-resolution phylogeographic
599 studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. *Molecular*
600 *Ecology Resources* **13**:929–937. DOI: 10.1111/1755-0998.12128.

Amiryousefi A, Hyvönen J, Poczai P. 2018a. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. *PLoS ONE* **13**: 1–23. DOI: 10.1371/journal.pone.0196069.

Amiryousefi A, Hyvönen J, Poczai P. 2018b. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* **34**: 3030–3031 DOI: 10.1093/bioinformatics/bty220.

Aoki S, Ito M. 2000. Molecular phylogeny of *Nicotiana* (Solanaceae) based on the nucleotide sequence of the *mat*K gene. *Plant Biology* **2**: 316–324. DOI: 10.1055/s-2000-3710.

Asaf S, Khan AL, Khan AR, Waqas M, Kang S-M, Khan MA, Lee S-M, Lee I-J. 2016. Complete Chloroplast Genome of *Nicotiana otophora* and its Comparison with Related Species. *Frontiers in Plant Science* **7**: 1–12. DOI: 10.3389/fpls.2016.00843.

Beier S, Thiel T, Münch T, Scholz U, Mascher M. 2017. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**: 2583–2585. DOI: 10.1093/bioinformatics/btx198.

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573–580. DOI: 10.1093/nar/27.2.573.

Cao J, Jiang D, Zhao Z, Yuan S, Zhang Y, Zhang T, Zhong W, Yuan Q, Huang L. 2018. Development of Chloroplast Genomic Resources in Chinese Yam (*Dioscorea polystachya*). *BioMed Research International*. DOI: 10.1155/2018/6293847.

Casano LM, Martín M, Sabater B. 2001. Hydrogen peroxide mediates the induction of chloroplastic *ndh* complex under photooxidative stress in Barley. *Plant Physiology* **125**: 1450–1458 DOI: 10.1104/pp.125.3.1450.

Chase MW, Knapp S, Cox A V., Clarkson JJ, Butsko Y, Joseph J, Savolainen V, Parokonny AS. 2003. Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). *Annals of Botany* **92**: 107–127 DOI: 10.1093/aob/mcg087.

Chen X, Cui Y, Nie L, Hu H, Xu Z, Sun W, Gao T, Song J, Yao H. 2019. Identification and Phylogenetic Analysis of the Complete Chloroplast Genomes of Three *Ephedra* Herbs Containing Ephedrine. *BioMed Research International* **2019**: 1–10. DOI: 10.1155/2019/5921725.

Cheng H, Li J, Zhang H, Cai B, Gao Z, Qiao Y, Mi L. 2017. The complete chloroplast genome sequence of strawberry (*Fragaria* × *ananassa* Duch.) and comparison with related species of Rosaceae. *PeerJ* **5**: e3919 DOI: 10.7717/peerj.3919.

Choi KS, Chung MG, Park S. 2016. The complete chloroplast genome sequences of three Veroniceae species (Plantaginaceae): comparative analysis and highly divergent regions. *Frontiers in Plant Science* **7**:1–8. DOI: 10.3389/fpls.2016.00355.

Cooper G. 2000. *Chloroplasts and other plastids in the cell: A molecular approach. 2nd edition*.

Couvreur TLP, Chatrou LW, Sosef MSM, Richardson JE. 2008. Molecular phylogenetics reveal multiple tertiary vicariance origins of the African rain forest trees. *BMC Biology* **6**: 54.

Daniell H. 2007. Transgene containment by maternal inheritance: Effective or elusive? *Proceedings of the National Academy of Sciences of the USA* **104**: 6879–6880 DOI: 10.1073/pnas.0702219104.

Daniell H, Lin C-S, Yu M, Chang W-J. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* **17**:134. DOI: 10.1186/s13059-016-1004-2.

Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**: 2455–2457 DOI: 10.1093/bioinformatics/btq429.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M, Claverie J-M, Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**: W465–W469. DOI: 10.1093/nar/gkn180.

Dillon MO, Tu T, Xie L, Quipuscia Silvestre V. 2009. Biogeographic diversification in *Nolana* (Solanaceae), a ubiquitous member of the Atacama and Peruvian Deserts along the western coast of South America. *Journal of Systematics and Evolution* **47**: 457–476

Doan TM. 2003. A south-to-north biogeographic hypothesis for Andean speciation: evidence from the lizard genus *Proctoporus* (Reptilia, Gymnophthalmidae). *Journal of Biogeography* **30**: 361–374

659 **Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006.** Relaxed phylogenetics and dating
660 with confidence. *PLoS Biology* **4**: e88

661 **Drummond AJ, Suchard M, Xie D, Rambaut A. 2012.** Bayesian phylogenetics with BEAUti
662 and the BEAST 1.7. *Molecular Biology and Evolution* **29**:1969 – 1973

663 **Erixon P, Oxelman B. 2008.** Whole-gene positive selection, elevated synonymous substitution
664 rates, duplication, and indel evolution of the chloroplast *clp*P1 gene. *PLoS ONE* **3**: e1386
665 DOI: 10.1371/journal.pone.0001386.

666 **Evans M, Aubriot X, Hearn D, Lanciaux M, Lavergne S, Cruaud C, Lowry II PP,**
667 **Haevermans R. 2014.** The evolution of succulence: insights from a remarkable radiation in
668 Madagascar. *Systematic Biology* **63**: 698 – 711.

669 **Garreaud RD, Molina A, Farias M. 2010**. Adean uplift ocean cooling and Atacama
670 hyeraridity: a climate modeling perspective. *Earth and Planetary Science Letters* **292**: 39–
671 50

672 **Gatenby AA, Rothstein SJ, Nomura M. 1989.** Translational coupling of the maize chloroplast
673 *atp*B and *atp*E genes. *Proceedings of the National Acadademy of Sciences of the USA* **86**:
674 4066–4070

675 **Goodspeed TH. 1954.** The Genus *Nicotiana*. Chronica Botanica, New York, USA**Greiner S,**
676 **Lehwark P, Bock R. 2019.** OrganellarGenomeDRAW (OGDRAW) version 1.3.1:
677 expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids*
678 *Research* 47: W59–W64 DOI: 10.1093/nar/gkz238.

679 **Hajdukiewicz PTJ, Allison LA, Maliga P. 1997.** The two RNA polymerases encoded by the
680 nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids.
681 *EMBO Journal* **16**: 4041–4048 DOI: 10.1093/emboj/16.13.4041.

682 **Hao DC, Chen SL, Xiao PG. 2010.** Molecular evolution and positive Darwinian selection of the
683 chloroplast maturase *mat*K. *Journal of Plant Research* **123**: 241–247 DOI: 10.1007/s10265-
684 009-0261-5.

685 **Hausner G, Olson R, Simon D, Johnson I, Sanders ER, Karol KG, McCourt RM,**
686 **Zimmerly S. 2006.** Origin and evolution of the chloroplast *trn*K (*mat*K) intron: A model for
687 evolution of group II intron RNA structures. *Molecular Biology and Evolution* **23**: 380–391

DOI: 10.1093/molbev/msj047.

Henriquez CL, Arias T, Pires JC, Croat TB, Schaal BA. 2014. Phylogenomics of the plant family Araceae. *Molecular Phylogenetics and Evolution* **75**: 91–102. DOI: 10.1016/j.ympev.2014.02.017.

Hilu KW, Alice LA. 2001. A phylogeny of Chloridoideae (Poaceae) based on *matK* sequences. *Systematic Botany* **26**: 386–405

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**: 518–522

Hoorn C, Wesselingh FP, ter Steege H, Bermudez MA, Mora A, Sevink J, et al. 2010. Amazonia through tie: Andean uplift, climate change, landscape evolution, and biodiversity. *Science* **330**: 927–931

Horvath EM, Peter SO, Joet T, Rumeau D, Cournac L, Horvath G V., Kavanagh TA, Schafer C, Peltier G, Medgyesy P. 2000. Targeted inactivation of the plastid ndhB gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. *Plant Physiology* **123**: 1337–1350

Jheng C-F, Chen T-C, Lin J-Y, Chen T-C, Wu W-L, Chang C-C. 2012. The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis orchids*. *Plant Science* **190**: 62–73. DOI: 10.1016/j.plantsci.2012.04.001.

Jin S, Daniell H. 2015. Engineered chloroplast genome just got smarter. *Trends in Plant Science* 20:622-640 DOI: 10.1016/j.tplants.2015.07.004.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589. DOI: 10.1038/nmeth.4285.

Kapralov M V., Kubien DS, Andersson I, Filatov DA. 2011. Changes in Rubisco kinetics during the evolution of C4 Photosynthesis in *Flaveria* (Asteraceae) are associated with positive selection on genes encoding the enzyme. *Molecular Biology and Evolution* **28**: 1491–1503. DOI: 10.1093/molbev/msq335.

716 **Katoh K, Kuma KI, Toh H, Miyata T. 2005.** MAFFT version 5: Improvement in accuracy of
717 multiple sequence alignment. *Nucleic Acids Research* **33**: 511–518. DOI:
718 10.1093/nar/gki198.

719 **Katoh K, Standley DM. 2013.** MAFFT multiple sequence alignment software version 7:
720 Improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.
721 DOI: 10.1093/molbev/mst010.

722 **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A,**
723 **Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012.**
724 Geneious Basic: an integrated and extendable desktop software platform for the
725 organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649. DOI:
726 10.1093/bioinformatics/bts199.

727 **Kimura M. 1979.** Model of effectively neutral mutations in which selective constraint is
728 incorporated. *Proceedings of the National Academy of Sciences of the USA* **76**: 3440–3444
729 DOI: 10.1073/pnas.76.7.3440.

730 **Knapp S, Bohs L, Nee M, Spooner DM. 2004.** Solanaceae - A model for linking genomics with
731 biodiversity. *Comparative and Functional Genomics* **5**: 285–291. DOI: 10.1002/cfg.393.

732 **Knapp S, Chase MW, Clarkson JJ. 2004.** Nomenclatural changes and a new sectional
733 classification in *Nicotiana* (Solanaceae). *Taxon* **53**: 73-82. DOI: 10.2307/4135490.

734 **Kumar S, Strecher G, Suleski M, Hedges SB. 2017.** TimeTree: a resource for timelines,
735 timetrees, and divergence times. *Molecular Biology and Evolution* **34**: 1812–1819.

736 **Kurtz S. 2002.** REPuter: the manifold applications of repeat analysis on a genomic scale.
737 *Nucleic Acids Research* **29**: 4633–4642. DOI: 10.1093/nar/29.22.4633.

738 **Kusumi K, Sakata C, Nakamura T, Kawasaki S, Yoshimura A, Iba K. 2011.** A plastid
739 protein NUS1 is essential for build-up of the genetic system for early chloroplast
740 development under cold stress conditions. *Plant Journal* **68**: 1039–1050. DOI:
741 10.1111/j.1365-313X.2011.04755.x.

742 **Laslett D, Canback B. 2004.** ARAGORN, a program to detect tRNA genes and tmRNA genes
743 in nucleotide sequences. *Nucleic Acids Research* **32**: 11–16. DOI: 10.1093/nar/gkh152.

744 **Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013.** Strong Purifying Selection at
745 Synonymous Sites in *D. melanogaster*. *PLoS Genetics* **9**: e1003527. DOI:
746 10.1371/journal.pgen.1003527.

747 **Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR. 2008.** The
748 ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae).
749 *Annals of Botany* **101**: 805–814. DOI: 10.1093/aob/mcm326.

750 **Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S,**
751 **Gascuel O. 2019.** NGPhylogeny.fr: new generation phylogenetic services for non-
752 specialists. *Nucleic Acids Research* **47**: W260–W265. DOI: 10.1093/nar/gkz303.

753 **Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler transform.
754 *Bioinformatics* **25**:1754–1760. DOI: 10.1093/bioinformatics/btp324.

755 **Li Y, Zhang Z, Yang J, Lv G. 2018.** Complete chloroplast genome of seven *Fritillaria* species,
756 variable DNA markers identification and phylogenetic relationships within the genus. *PLoS*
757 *ONE* **13**: e0194613 DOI: 10.1371/journal.pone.0194613.

758 **Liu L, Wang Y, He P, Li P, Lee J, Soltis DE, Fu C. 2018.** Chloroplast genome analyses and
759 genomic resource development for epilithic sister genera *Oresitrophe* and *Mukdenia*
760 (Saxifragaceae), using genome skimming data. *BMC Genomics* **19**: 1–17. DOI:
761 10.1186/s12864-018-4633-x.

762 **Lowe TM, Chan PP. 2016.** tRNAscan-SE On-line: integrating search and context for analysis of
763 transfer RNA genes. *Nucleic Acids Research* **44**: W54–W57 DOI: 10.1093/nar/gkw413.

764 **Luebert F, Weigend M. 2014.** Phylogenetic insights into Andean plant diversification.
765 *Frontiers in Ecology and Evolution* **2**: 27

766 **Majure LC, Puente R, Patrick Griffith M, Judd WS, Soltis PS, Soltis DE. 2012.** Phylogeny
767 of *Opuntia* s.s. (Cactaceae): Clade delineation, geographic origins, reticulate evolution.
768 *American Journal of Botany* **99**: 847–864 DOI: 10.3732/ajb.1100375.

769 **Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B,**
770 **Hasegawa M, Penny D. 2002.** Evolutionary analysis of *Arabidopsis*, cyanobacterial, and
771 chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the
772 nucleus. *Proceedings of the National Academy of Sciences of the USA* **99**: 12246–12251.

DOI: 10.1073/pnas.182432999.

Mehmood F, Abdullah, Shahzadi I, Ahmed I, Waheed MT, Mirza B. 2019. Characterization of *Withania somnifera* chloroplast genome and its comparison with other selected species of Solanaceae. *Genomics*. DOI: 10.1016/J.YGENO.2019.08.024.

Menezes APA, Resende-Moreira LC, Buzatti RSO, Nazareno AG, Carlsen M, Lobo FP, Kalapothakis E, Lovato MB. 2018. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. *Scientific Reports* **8**: 1–12. DOI: 10.1038/s41598-018-20189-4.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. 2009. Tablet-next generation sequence assembly visualization. *Bioinformatics* **26**: 401–402. DOI: 10.1093/bioinformatics/btp666.

Mower JP. 2009. The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Research* **37**: W253–W259. DOI: 10.1093/nar/gkp337.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Molecular Biology and Evolution* **30**: 1196-1205. DOI: 10.1093/molbev/mst030.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* **8**: e1002764. DOI: 10.1371/journal.pgen.1002764.

Muse S V, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**: 715–724. DOI: 10.1093/oxfordjournals.molbev.a040152.

Nakajima Y, Mulligan RM. 2001. Heat stress results in incomplete C-to-U edting of maize chloroplast mRNAs and correlates with changes in chloroplast transcription rate. *Current Genetics* **40**: 209–213. DOI: 10.1007/s002940100249.

Neale DB, Sederoff RR. 1989. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. *Theoretical and Applied Genetics* **77**:

212–216. DOI: 10.1007/BF00266189.

Nguyen VB, Park H-S, Lee S-C, Lee J, Park JY, Yang T-J. 2017. Authentication markers for five major *Panax* species developed via comparative analysis of complete chloroplast genome sequences. *Journal of Agricultural and Food Chemistry* **65**: 6298–6306 DOI: 10.1021/acs.jafc.7b00925.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating Maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274. DOI: 10.1093/molbev/msu300.

Occhialini A, Lin MT, Andralojc PJ, Hanson MR, Parry MAJ. 2016. Transgenic tobacco plants with improved cyanobacterial Rubisco expression but no extra assembly factors grow at near wild-type rates if provided with elevated CO2. *Plant Journal* **85**: 148–160. DOI: 10.1111/tpj.13098.

Oldenburg DJ, Bendich AJ. 2015. DNA maintenance in plastids and mitochondria of plants. *Frontiers in Plant Science* **6**: 883. DOI: 10.3389/fpls.2015.00883.

Oldenburg DJ, Bendich AJ. 2016. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. *Current Genetics* **62**: 431–442. DOI: 10.1007/s00294-015-0548-0.

Olmstead RG, Bohs L. 2007. A summary of molecular systematic research in solanaceae: 1982-2006. *Acta Horticulturae* **745**: 255–268. DOI: 10.17660/ActaHortic.2007.745.11.

Olmstead RG, Bohs L, Migid HA, Santiago-valentin E. 2008. A molecular phylogeny of the Solanaceae. *Taxon* **57**: 1159–1181

Palmer JD. 1985. Comparative organization of chloroplast genomes. *Annual Review of Genetics* **19**: 325–354. DOI: 10.1146/annurev.ge.19.120185.001545.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology* 14: e1002379

Peng L, Shikanai T. 2011. Supercomplex formation with photosystem I is required for the stabilization of the chloroplast NADH dehydrogenase-like complex in *Arabidopsis*. *Plant Physiology* **155**: 1629–1639. DOI: 10.1104/pp.110.171264.

Picard D, Sempere T, Plantard O. 2008. Direction and timing of uplift propagation in the Peruvian Andes deduced from molecular phylogenetics of highland biotaxa. *Earth and Planetary Science Letters* **271**: 326–336

Piot A, Hackel J, Christin PA, Besnard G. 2018. One-third of the plastid genes evolved under positive selection in PACMAD grasses. *Planta* **247**: 255–266. DOI: 10.1007/s00425-017-2781-x.

Poczai P, Hyvönen J, Symon DE. 2011. Phylogeny of kangaroo apples (*Solanum* subg. *Archaesolanum*, Solanaceae). *Molecular Biology Reports* **38**: 5243–5259. DOI: 10.1007/s11033-011-0675-8.

Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, Yao H, Sun C, Li X, Li C, Liu J, Xu H, Chen S. 2013. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PloS ONE* **8**: e57607. DOI: 10.1371/journal.pone.0057607.

Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer v1.6. Computer program and documentation distributed by the author, website: http://beast.community/tracer . [accessed 3 December 2019].

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: multiple alignment of coding SEquences Accounting for frameshifts and stop codons. *PLoS One* **6**: e22594

Ravi V, Khurana JP, Tyagi AK, Khurana P. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* **271**: 101–122. DOI: 10.1007/s00606-007-0608-0.

Rodrigues NF, Christoff AP, da Fonseca GC, Kulcheski FR, Margis R. 2017. Unveiling chloroplast RNA editing events using next generation small RNA sequencing data. *Frontiers in Plant Science* **8**: 1686. DOI: 10.3389/fpls.2017.01686.

Ronquist F. 2004. Bayesian inference of character evolution. *Trends in Ecology and Evolution* **19**: 475–481

Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution* **34**: 3299–3302. DOI: 10.1093/molbev/msx248.

Rumeau D, Peltier G, Cournac L. 2007. Chlororespiration and cyclic electron flow around PSI

858 during photosynthesis and plant stress response. *Plant, Cell and Environment* **30**: 1041–
859 1051. DOI: 10.1111/j.1365-3040.2007.01675.x.

860 **Saina JK, Li ZZ, Gichira AW, Liao YY. 2018.** The complete chloroplast genome sequence of
861 tree of heaven (*Ailanthus altissima* (mill.) (sapindales: Simaroubaceae), an important
862 pantropical tree. *International Journal of Molecular Sciences* **19**: E929 DOI:
863 10.3390/ijms19040929.

864 **Shahzadi I, Abdullah, Mehmood F, Ali Z, Ahmed I, Mirza B. 2019.** Chloroplast genome
865 sequences of *Artemisia maritima* and *Artemisia absinthium*: Comparative analyses,
866 mutational hotspots in genus *Artemisia* and phylogeny in family Asteraceae. *Genomics*.
867 DOI: 10.1016/J.YGENO.2019.08.016.

868 **Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C. 2019.** CPGAVAS2, an integrated
869 plastome sequence annotator and analyzer. *Nucleic Acids Research* **47**: W65–W73. DOI:
870 10.1093/nar/gkz345.

871 **Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N,**
872 **Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY,**
873 **Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A,**
874 **Tohdoh N, Shimada H, Sugiura M. 1986.** The complete nucleotide sequence of the
875 tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal* **5**:
876 2043–2049. DOI: 10.1002/j.1460-2075.1986.tb04464.x.

877 **Sierro N, Battey JND, Bovet L, Liedschulte V, Ouadi S, Thomas J, Broye H, Laparra H,**
878 **Vuarnoz A, Lang G, Goepfert S, Peitsch MC, Ivanov N V. 2018.** The impact of genome
879 evolution on the allotetraploid *Nicotiana rustica* - An intriguing story of enhanced alkaloid
880 production. *BMC Genomics* **19**: 855 DOI: 10.1186/s12864-018-5241-5.

881 **Simpson, BB. 1975.** Pleistocene changes in the Flora of the high tropical Andes. *Paleobiology* **1**:
882 273–294

883 **Smith, H.H. 1974.** *Nicotiana*. In: King RC (eds) Handbook of Genetics 2. Plenum Press, New
884 York, pp 281-314

885 **Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M. 2005.** The
886 complete nucleotide sequence and multipartite organization of the tobacco mitochondrial

887 genome: Comparative analysis of mitochondrial genomes in higher plants. *Molecular*
888 *Genetics and Genomics* **272**: 603–615. DOI: 10.1007/s00438-004-1075-8.

889 **Terashima I, Araya T, Miyazawa SI, Sone K, Yano S. 2005.** Construction and maintenance of
890 the optimal photosynthetic systems of the leaf, herbaceous plant and tree: An eco-
891 developmental treatise. *Annals of Botany* **95**: 507–519. DOI: 10.1093/aob/mci049.

892 **Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017.**
893 GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*
894 **45**: W6–W11. DOI: 10.1093/nar/gkx391.

895 **Waheed MT, Ismail H, Gottschamel J, Mirza B, Lössl AG. 2015.** Plastids: the green frontiers
896 for vaccine production. *Frontiers in Plant Science* **6**: 1005. DOI: 10.3389/fpls.2015.01005.

897 **Waheed MT, Thönes N, Müller M, Hassan SW, Razavi NM, Lössl E, Kaul HP, Lössl AG.**
898 2011. Transplastomic expression of a modified human papillomavirus L1 protein leading to
899 the assembly of capsomeres in tobacco: A step towards cost-effective second-generation
900 vaccines. *Transgenic Research* **20**: 271–281. DOI: 10.1007/s11248-010-9415-4.

901 **Wambugu PW, Brozynska M, Furtado A, Waters DL, Henry RJ. 2015.** Relationships of
902 wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast
903 genome sequences. *Scientific Reports* **5**:13957. DOI: 10.1038/srep13957.

904 **Xu J, Feng D, Song G, Wei X, Chen L, Wu X, Li X, Zhu Z. 2003.** The first intron of rice
905 EPSP synthase enhances expression of foreign gene. *Science in China. Series C, Life*
906 *sciences / Chinese Academy of Sciences* **46**: 561–569. DOI: 10.1360/02yc0120.

907 **Xu J-H, Liu Q, Hu W, Wang T, Xue Q, Messing J. 2015.** Dynamics of chloroplast genomes in
908 green plants. *Genomics* **106**: 221–231. DOI: 10.1016/J.YGENO.2015.07.004.

909 **Yamori W. 2016.** Photosynthetic response to fluctuating environments and photoprotective
910 strategies under abiotic stress. *Journal of Plant Research* **129**: 379–395. DOI:
911 10.1007/s10265-016-0816-1.

912 **Yamori W, Shikanai T. 2016.** Physiological functions of cyclic electron transport around
913 photosystem I in sustaining photosynthesis and plant growth. *Annual Review of Plant*
914 *Biology* **67**: 81–106. DOI: 10.1146/annurev-arplant-043015-112002.

Yang Z, Wang G, Ma Q, Ma W, Liang L, Zhao T. 2019. The complete chloroplast genomes of three Betulaceae species: Implications for molecular phylogeny and historical biogeography. *PeerJ* **7**: e6320. DOI: 10.7717/peerj.6320.

Yu Y, Blair C, He XJ. 2019. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. *Molecular Biology and Evolution* doi: 10.1093/molbev/msz257

Yukawa M, Tsudzuki T, Sugiura M. 2006. The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: Complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. *Molecular Genetics and Genomics* **275**: 367–373. DOI: 10.1007/s00438-005-0092-6.

Zhang Y, Cui YL, Zhang XL, Yu QB, Wang X, Yuan XB, Qin XM, He XF, Huang C, Yang ZN. 2018. A nuclear-encoded protein, mTERF6, mediates transcription termination of *rpo*A polycistron for plastid-encoded RNA polymerase-dependent chloroplast gene expression and chloroplast development. *Scientific Reports* **8**: 11929. DOI: 10.1038/s41598-018-30166-6.

Zhang Y, Du L, Liu A, Chen J, Wu L, Hu W, Zhang W, Kim K, Lee S-C, Yang T-J, Wang Y. 2016. The Complete chloroplast genome sequences of five *Epimedium* species: lights into phylogenetic and taxonomic analyses. *Frontiers in Plant Science* **7**: 306. DOI: 10.3389/fpls.2016.00306.

Zhang J, Zhang Y, Du Y, Chen S, Tang H. 2011. Dynamic metabonomic responses of tobacco (*Nicotiana tabacum*) plants to salt stress. *Journal of Proteome Research* **10**: 1904–1914. DOI: 10.1021/pr101140n.

Zhao Z, Wang X, Yu Y, Yuan S, Jiang D, Zhang Y, Zhang T, Zhong W, Yuan Q, Huang L. 2018. Complete chloroplast genome sequences of *Dioscorea*: Characterization, genomic resources, and phylogenetic analyses. *PeerJ* **6**: e6032. DOI: 10.7717/peerj.6032.

Figure 1. Chloroplast genome map of *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia* and *Nicotiana glauca.*

Genes that lie outside the circle are transcribed clockwise while the genes that transcribed counterclockwise are inside the circle. Different colors indicate the genes belonging to various

functional groups. GC and AT content of genome are plotted light grey and dark, respectively, in the inner circle. Large single copy (LSC), inverted repeat A (IRa), inverted repeat B (IRb) and small single copy (SSC) are shown in the circular diagram. Inverted repeat regions are highlighted with *cinderella* color.

Figure 2. (A) Comparison of amino acid groups in *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia, Nicotiana glauca.* (B) Comparison of amino acid frequency in *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia, Nicotiana glauca.*

Figure 3 Comparison of the border positions of LSC, SSC and IR among the five *Nicotiana* chloroplast genomes.

Positive strand transcribed genes are indicated under the line while the genes that are transcribed by negative strands are indicated above the line. Gene names are expressed in boxes, and the lengths of relative regions are showed above the boxes. The number of bp (base pairs) that are written with genes reveal the part of the genes that exists in the region of chloroplast or away from region of chloroplast i.e. bp written with *ycf1* indicate that sequences exist in that region of the plastid genome.

Figure 4. Comparison of microsatellite repeats among *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia, Nicotiana glauca. (A)* Indicate numbers of various types of microsatellites present in the plastid genome of *Nicotiana* species. *(B)* Distribution of SSRs in different regions of the plastid genome of *Nicotiana* species. *(C)* SSRs motifs distribution in different regions of the plastid genome of *Nicotiana* species.

Figure 5. (A). Indication of various kinds of oligonucleotide repeats exist in all *Nicotiana species* (B). Indicate repeats that exist range of size i.e. 30–35 indicate numbers of repeats within the size vary from 30 and 35. (C). Indicate number of repeats exist in separate areas of plastid genome. LSC: Large single copy, SSC: small single copy, IR: inverted repeat region, LSC/SSC: one copy of LSC and another in SSC, LSC/IR: one copy of LSC and another in SSC, IR/SSC: one copy of IR and another in SSC, LSC/SSC/IR: one copy of LSC, one in SSC and another in IR. (D). Indicate number of repeats in different regions of plastid genome. IGS: Intergenic spacer region, CDS: coding DNA sequences, Intron: intronic regions, IGS/Intron: one copy of intergenic spacer region and another in intronic regions. Intron/CDS: one copy intron region and

another in CDS regions. IGS/CDS: intergenic spacer region copy of repeat and one more in coding regions.

Figure 6. Comparison of tandem repeats among *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia, Nicotiana glauca. (A)* Number of tandem repeats in the chloroplast genome of *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia, Nicotiana glauca. (B)* Location and number of tandem repeats in the plastid genome of *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia, Nicotiana glauca. (C)* Number, size, distribution of tandem repeats across the plastid genome of *Nicotiana knightiana, Nicotiana rustica, Nicotiana paniculata, Nicotiana obtusifolia, Nicotiana glauca*

Figure 7. Nucleotide diversity of various regions of the chloroplast genome among *Nicotiana* species. The X-axis indicate the chloroplast regions and Y-axis indicate the nucleotide diversity.

Figure 8. Maximum likelihood (ML) tree was reconstructed based on seventy-five protein coding plastid genes of eleven *Nicotiana* species and *Solanum dulcamara* as an outgroup. Bootstrap support values are shown above or below the nodes.

Figure 9. Plastome phylogeny and biogeography of the tetraploid *Nicotiana rustica* and related species. A) Map showing the six biogeographic areas used to infer the biogeographic history of the *Nicotiana rustica* in South America. Arrows illustrate the dispersal events inferred from the biogeographic analysis. Geographical distribution for each terminal is indicated using the biogeographic regions subdivision. The most probable ancestral area is figured at each node of the phylogenetic tree. Pie-charts represent relative probabilities of ancestral states at each node. B) Node-calibrated Bayesian maximum clade credibility tree with 95% highest posterior density (HPD) interval for node ages presented as horizontal bars and mean values are displayed above each node. All nodes have PP $\geq$ 0.97 and BS $\geq$ 87%. Trace plot of the combined chains showing the sampled joint probability and the convergence of the chains.

Table 1. List of amino acid replacements and results of positive selection tests on codons underlying these replacements.

1004 **Table 2.** Comparison of substitution in *Nicotiana* species

1005 **Table 3.** Distribution of indels in *Nicotiana* chloroplast genome

1006 **Table 4.** Mutational hotspots among *Nicotiana* species

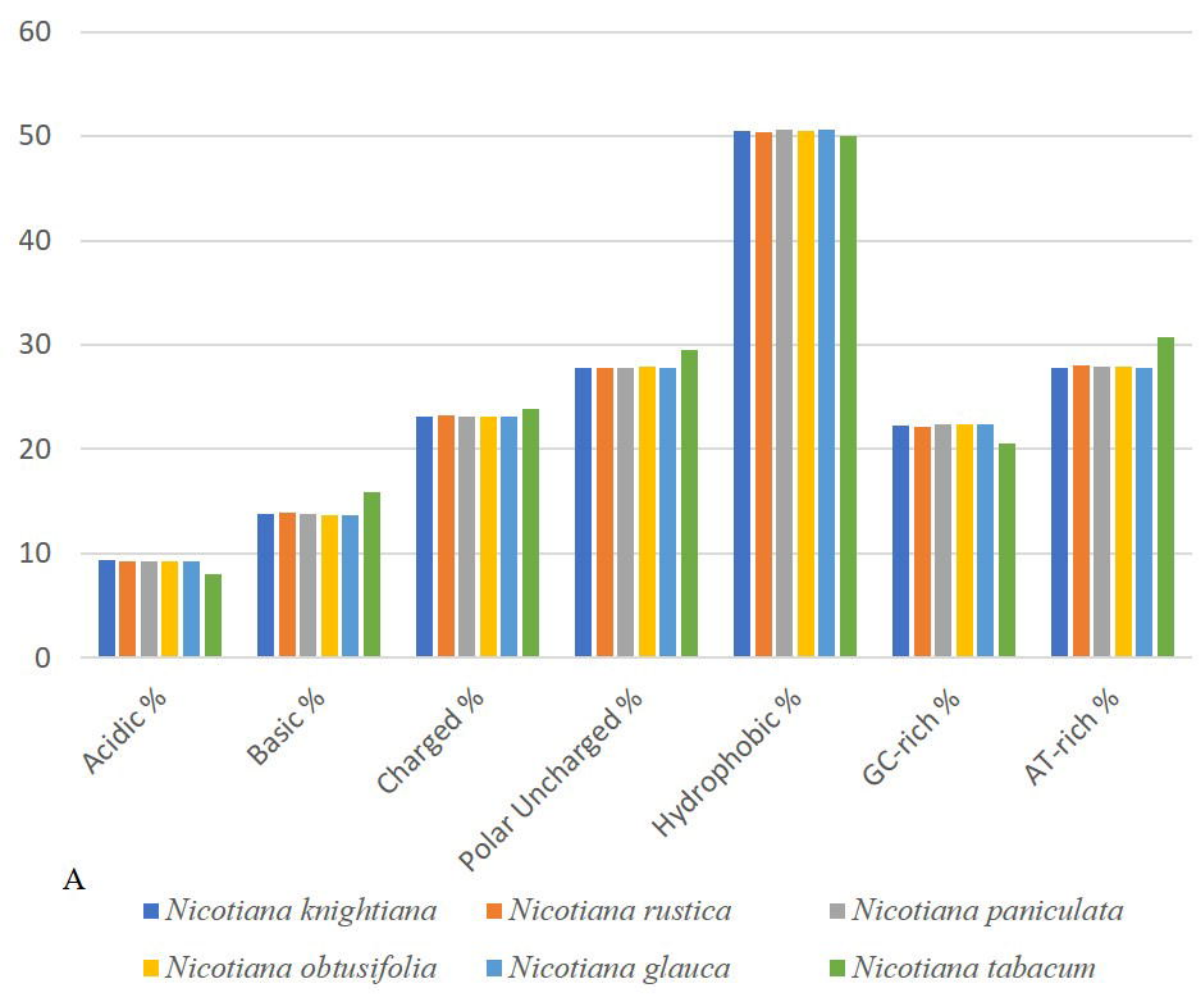
1007

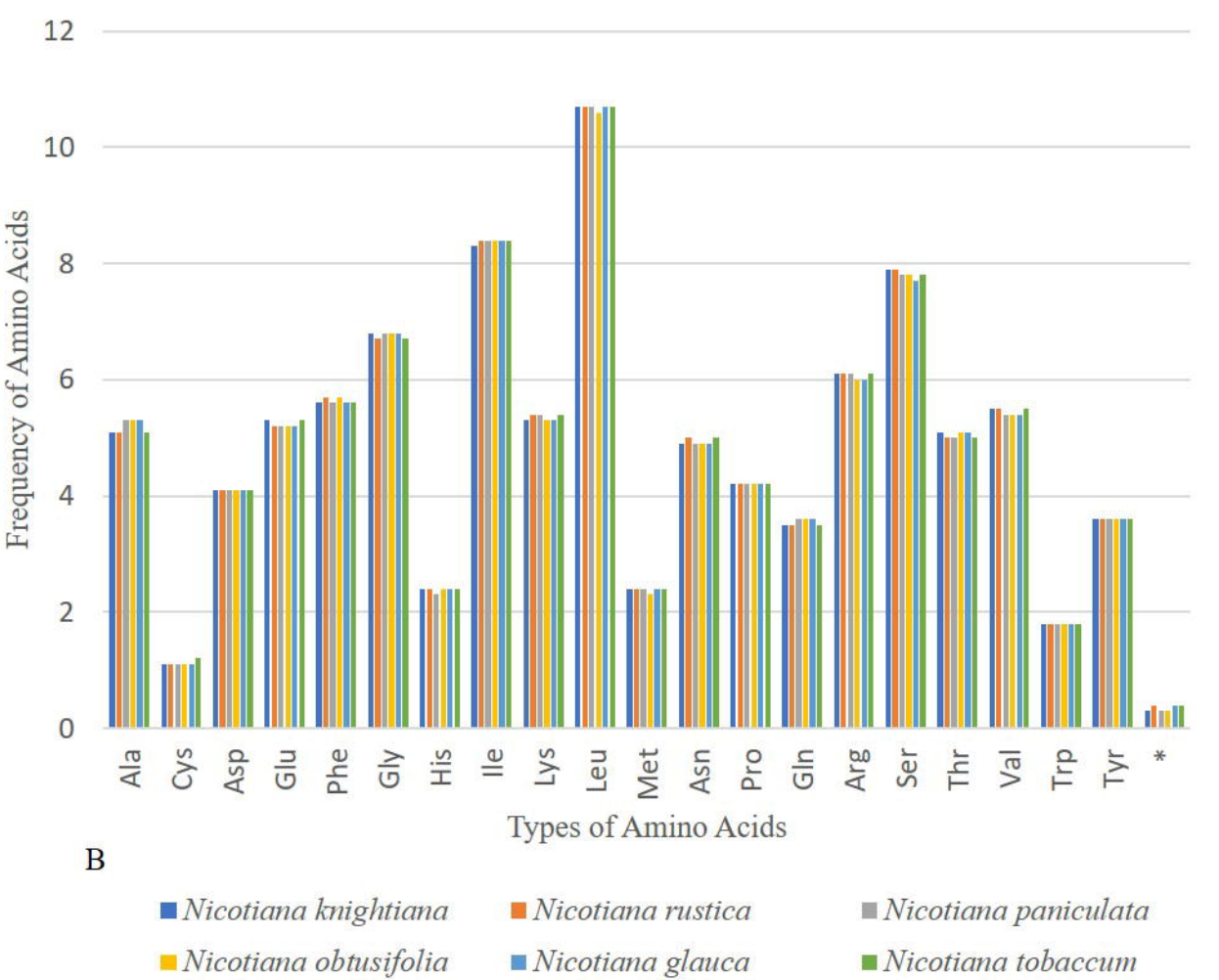
Nicotiana knightiana 155,969 bp
Nicotiana rustica 155,847 bp
Nicotiana paniculata 155,689 bp
Nicotiana obtusifolia 156,022 bp
Nicotiana glauca 155,917 bp

LSC
IRB
IRA
SSC

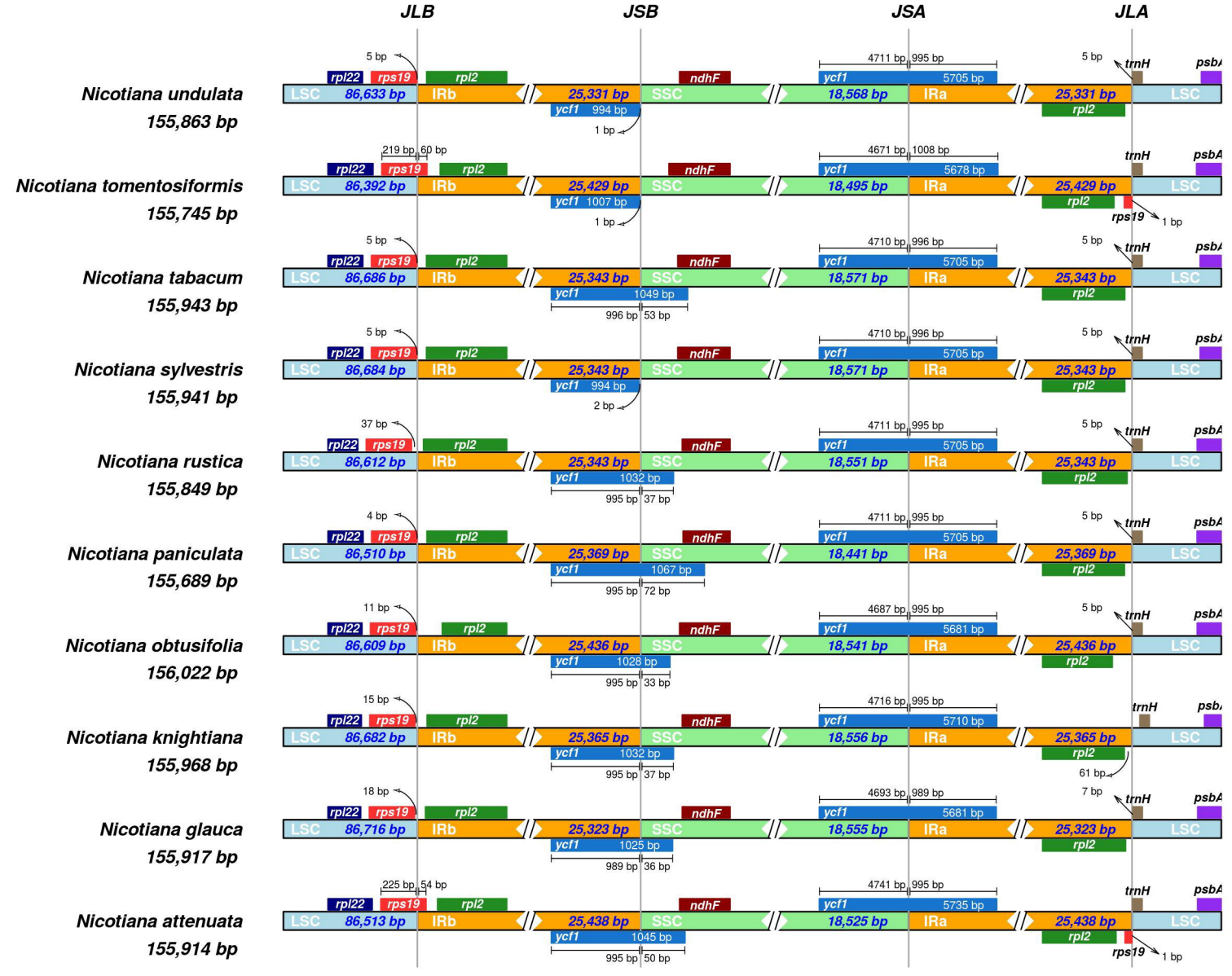
trnG*
psbZ
psbC
psbD
trnT
petN
trnC
rpoB
rpoC1*
rpoC2
rps2
atpI
atpH
atpF*
atpA
trnR
trnG*
trnS
psbI
psbK
trnQ
rps16*
trnK*
matK
psbA
trnH
rpl2*
rpl23
trnI*
ycf2
ORF79
trnL *
ndhB*
rps7
ORF70B
trnV*
rrn16
trnI*
trnA*
rrn23
rrn4.5
rrn5
trnR
ORF350; truncated ycf1
trnN
ORF75
ndhF
ndhD
psaC
ndhE
ndhG
ndhI
ndhA*
rps15
ndhH
ycf1
trnS
rps14
trnfM
psaA
psaB
ycf3*
trnS
trnT
rps4
trnF
trnL *
ndhJ
ndhK
ndhC
trnM
atpE
atpB
trnV*
rbcL
accD
psaI
ycf4
cemA
petA
petL
petG
psaJ
rpl33
rps18
rpl20
N/D
N/D
psbB
psbT
psbH
psbN
petB*
N/D
rpoA
rps11
rpl36
initiation factor
rps8
rpl14
rpl16*
rps3
rpl22
rps19
rpl2*
rpl23
trnI*
ycf2
trnL *
ndhB*
rps7
N/D
N/D
ORF131
ORF70B
trnV*
rrn16
trnI*
trnA*
rrn23
rrn4.5
rrn5
trnR
ORF350; truncated ycf1
trnN
ORF75
rpl32
sprA
trnL *
ccsA
psbJ
psbL
psbF
psbE
trnW
trnP

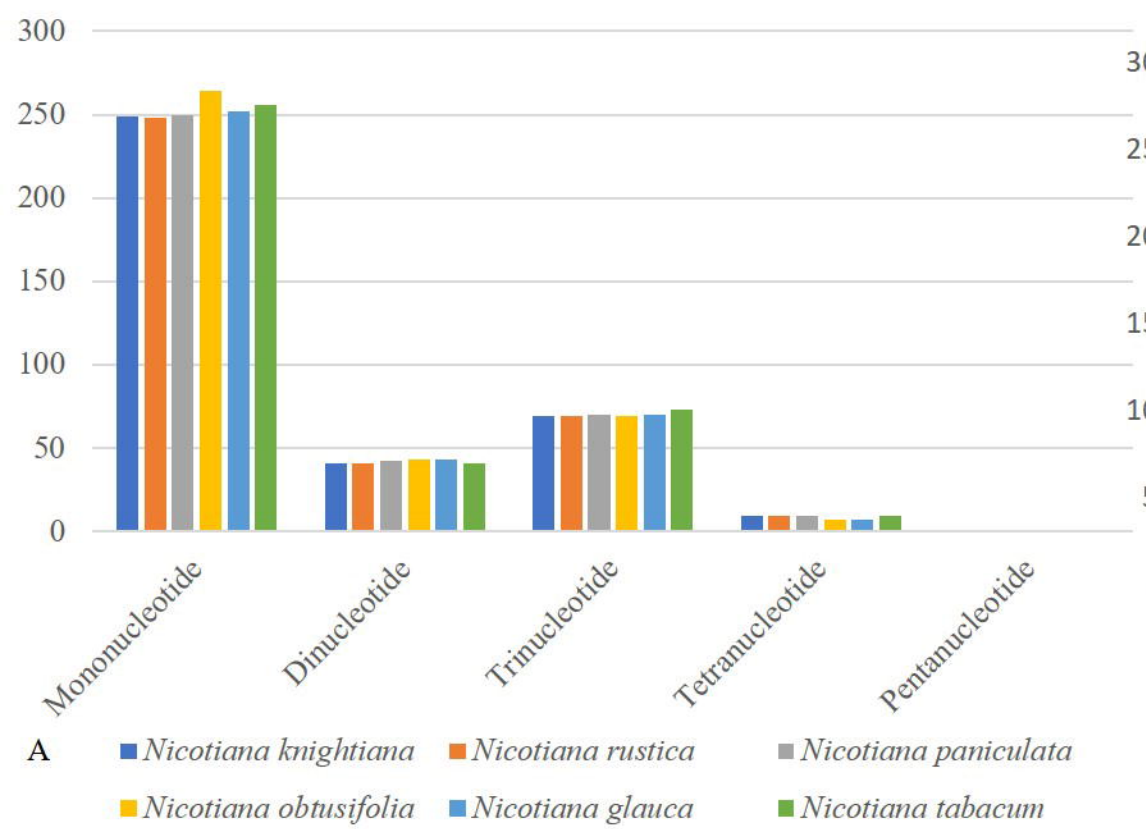
- photosystem I
- photosystem II
- cytochrome b/f complex
- ATP synthase
- NADH dehydrogenase
- RubisCO large subunit
- RNA polymerase
- ribosomal proteins (SSU)
- ribosomal proteins (LSU)
- clpP, matK
- other genes
- hypothetical chloroplast reading frames (ycf)
- ORFs
- transfer RNAs
- ribosomal RNAs
- origin of replication
- polycistronic transcripts


60
50
40
30
20
10
0
Acidic %
Basic %
Charged %
Polar Uncharged %
Hydrophobic %
GC-rich %
AT-rich %
A
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

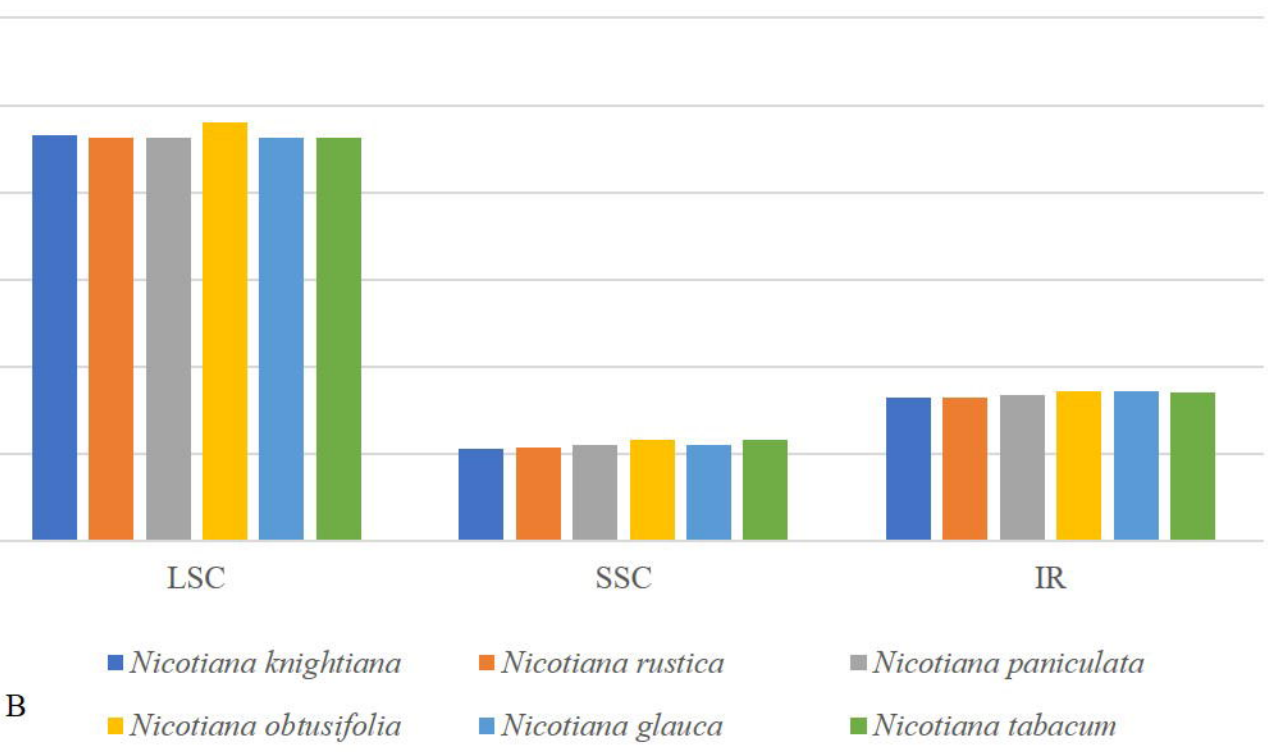

12
10
8
6
4
2
0
Frequency of Amino Acids
Ala
Cys
Asp
Glu
Phe
Gly
His
Ile
Lys
Leu
Met
Asn
Pro
Gln
Arg
Ser
Thr
Val
Trp
Tyr
*
Types of Amino Acids
B
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tobaccum

Inverted Repeats

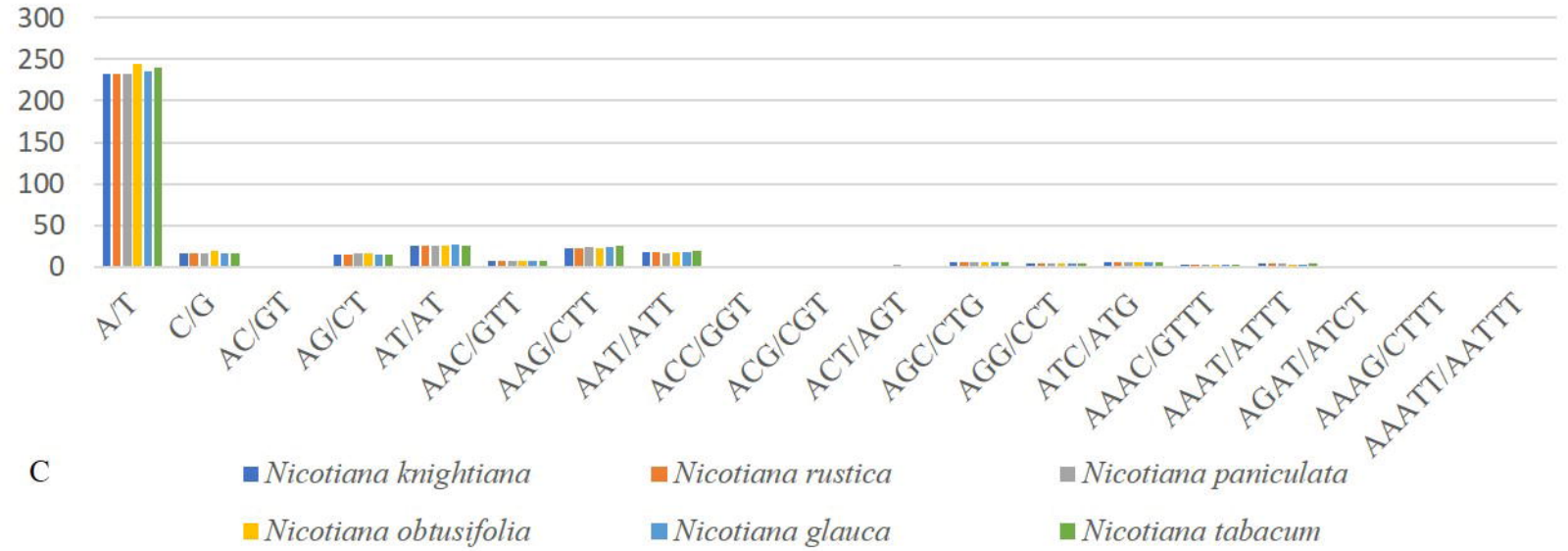



300
250
200
150
100
50
0
Mononucleotide
Dinucleotide
Trinucleotide
Tetranucleotide
Pentanucleotide
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

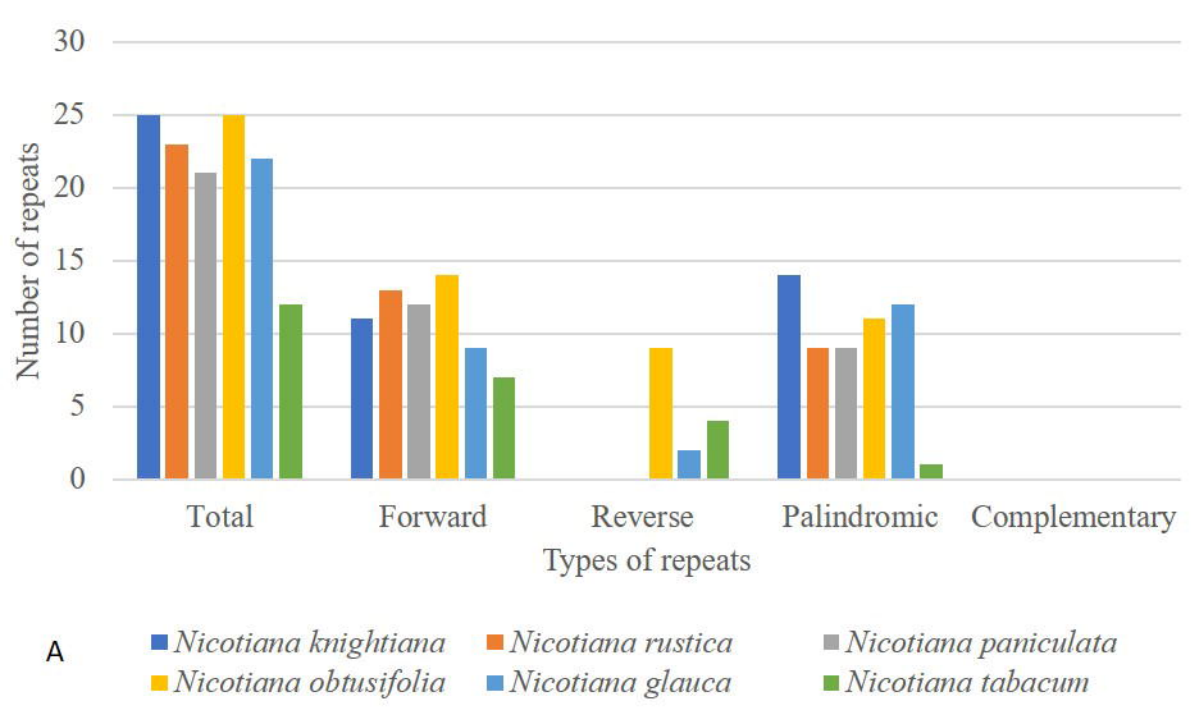
A


300
250
200
150
100
50
0
LSC
SSC
IR
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

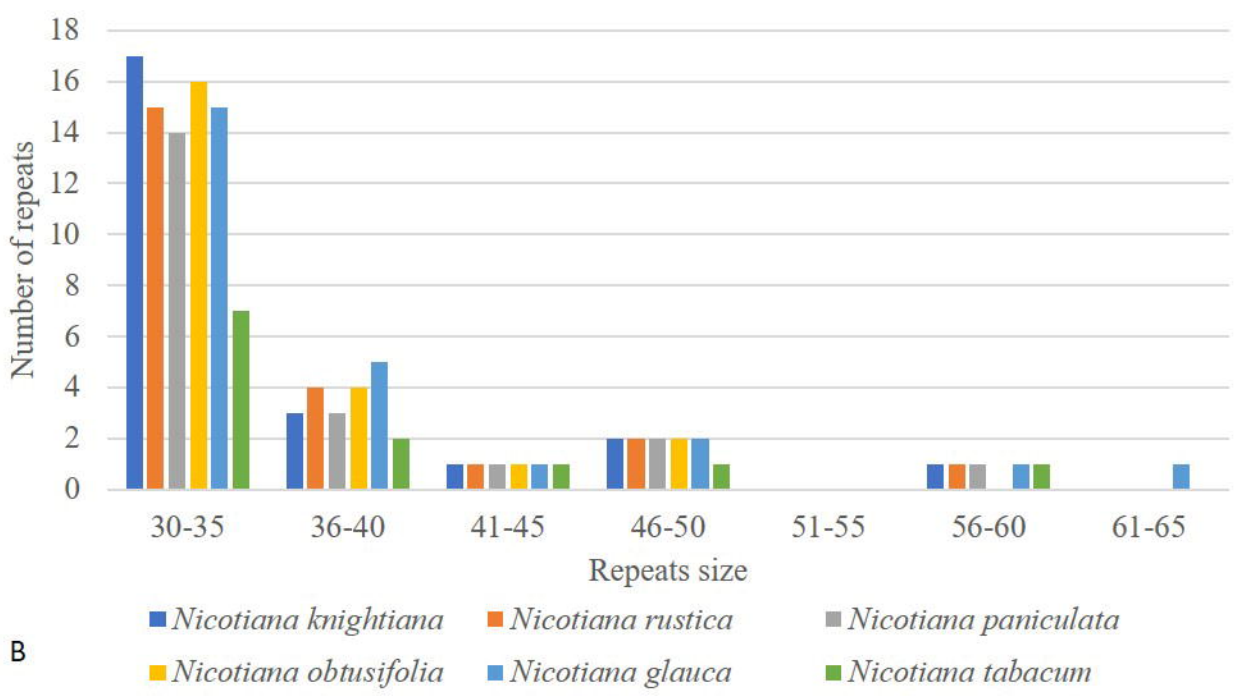
B


300
250
200
150
100
50
0
A/T
C/G
AC/GT
AG/CT
AT/AT
AAC/GTT
AAG/CTT
AAT/ATT
ACC/GGT
ACG/CGT
ACT/AGT
AGC/CTG
AGG/CCT
ATC/ATG
AAAC/GTTT
AAAT/ATTT
AGAT/ATCT
AAAG/CTTT
AAATT/AATTT
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

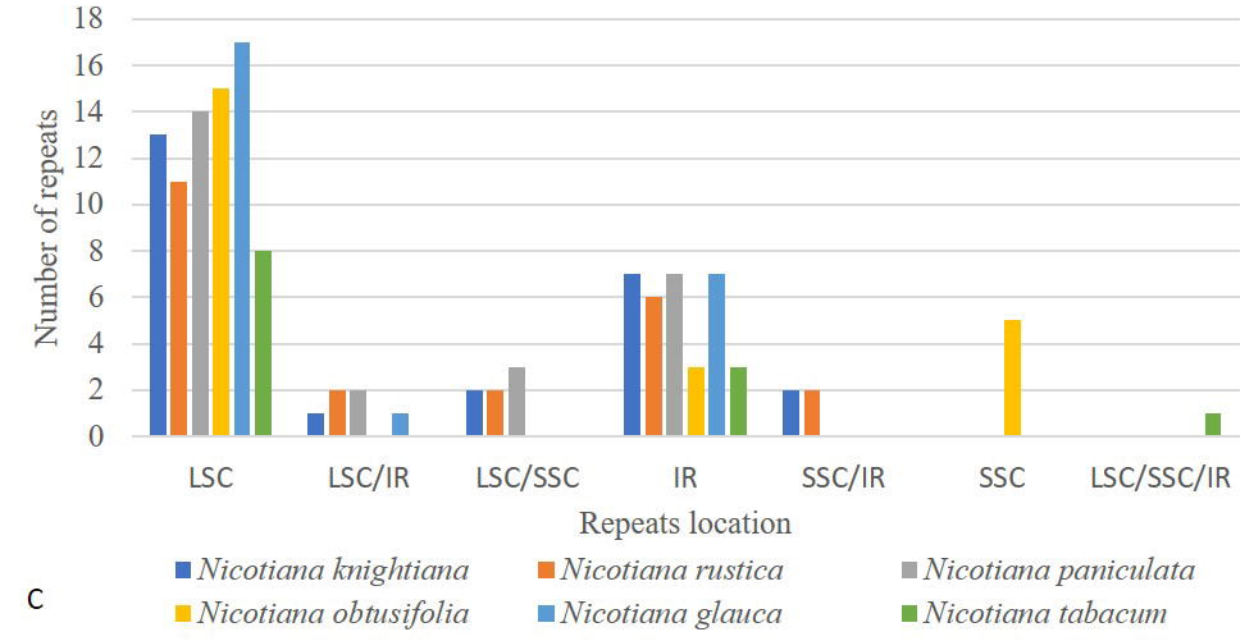
C


Number of repeats
30
25
20
15
10
5
0
Total
Forward
Reverse
Palindromic
Complementary
Types of repeats
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

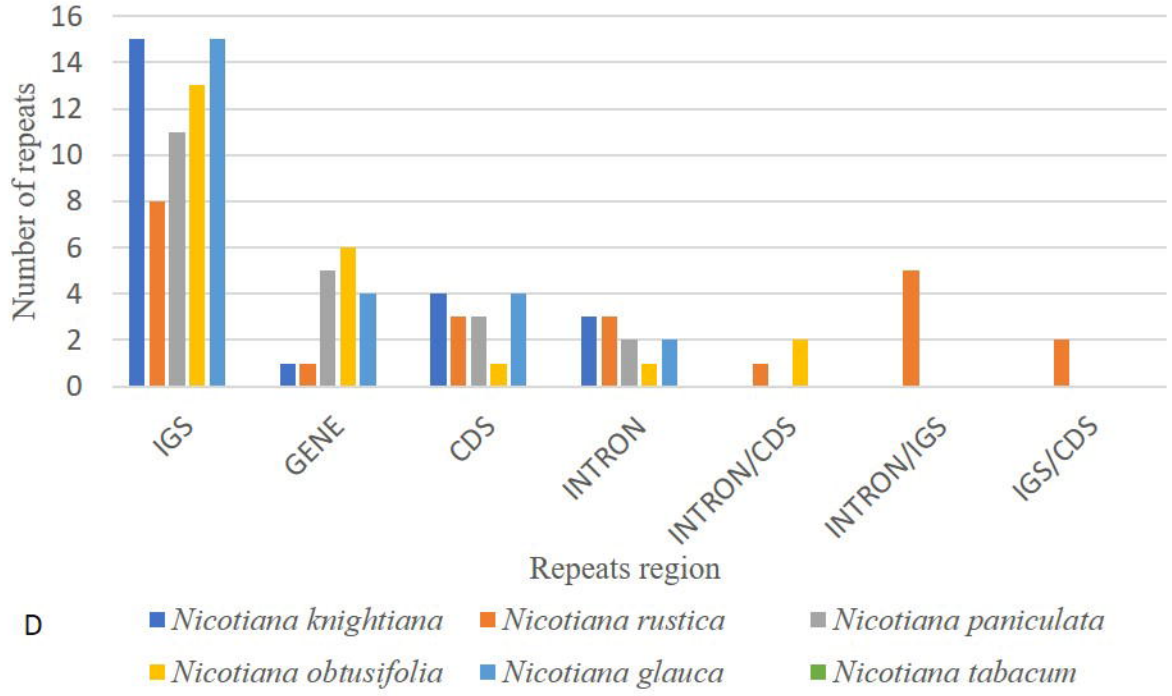
A


Number of repeats
18
16
14
12
10
8
6
4
2
0
30-35
36-40
41-45
46-50
51-55
56-60
61-65
Repeats size
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

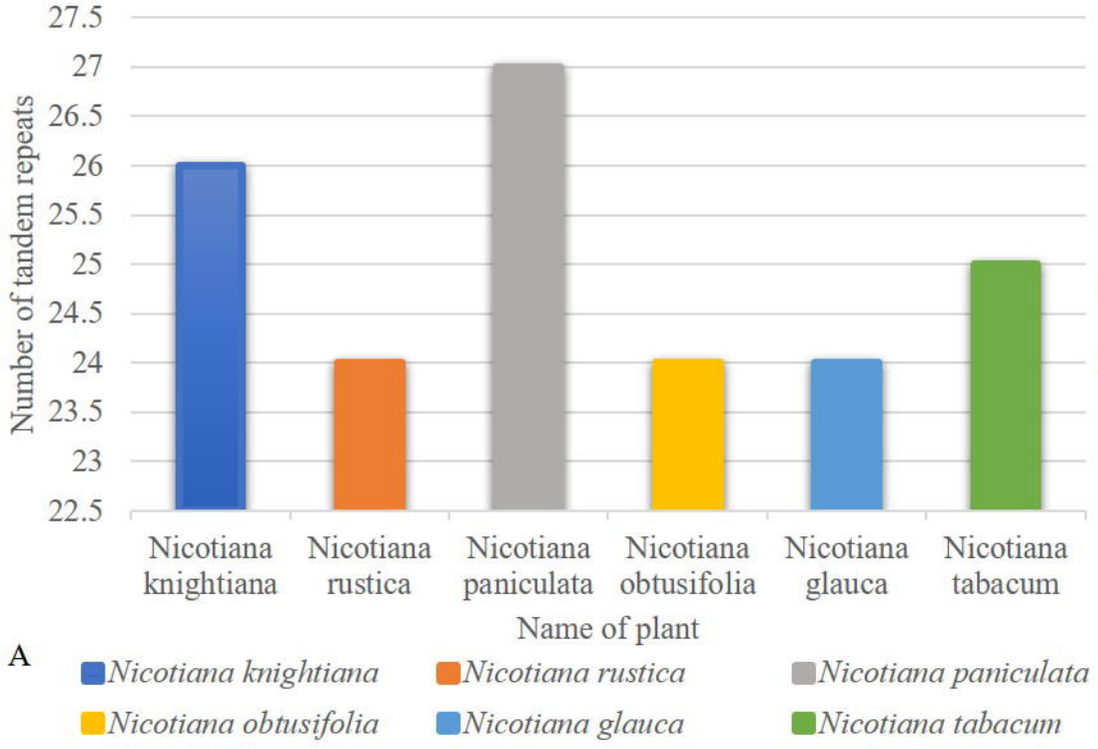
B

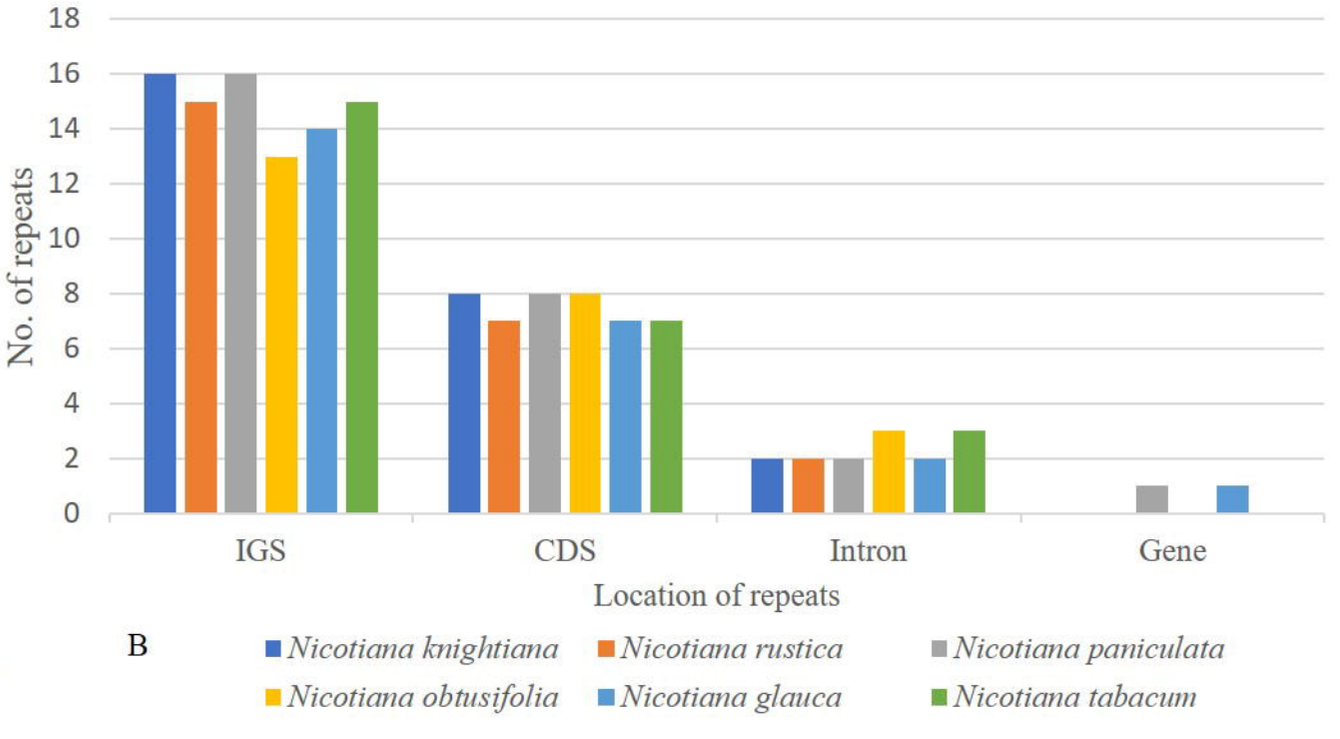

Number of repeats
18
16
14
12
10
8
6
4
2
0
LSC
LSC/IR
LSC/SSC
IR
SSC/IR
SSC
LSC/SSC/IR
Repeats location
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

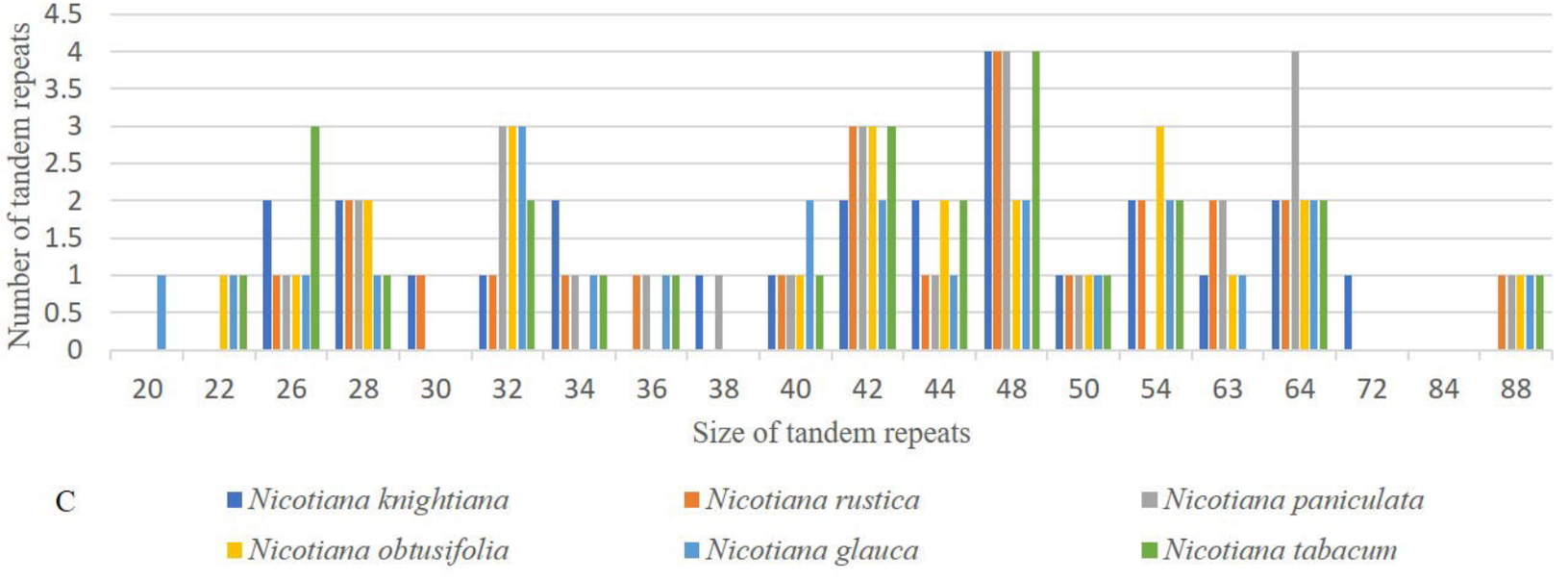
C

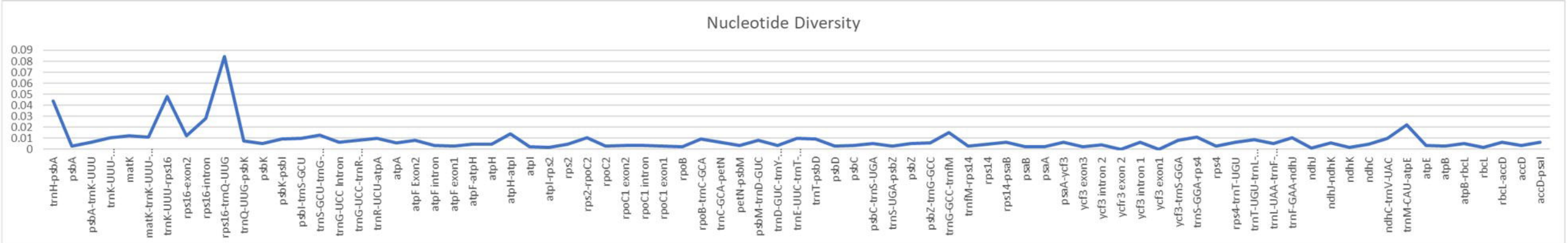

Number of repeats
16
14
12
10
8
6
4
2
0
IGS
GENE
CDS
INTRON
INTRON/CDS
INTRON/IGS
IGS/CDS
Repeats region
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum

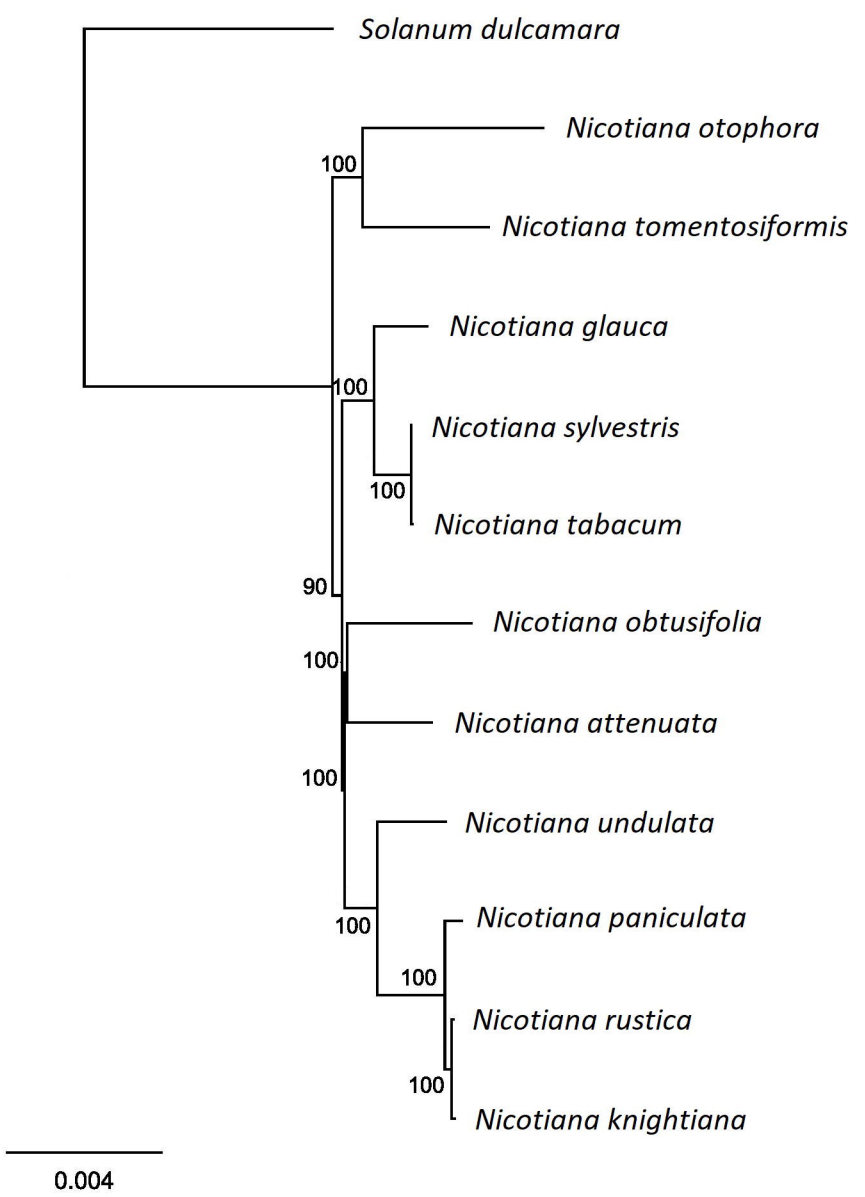
D

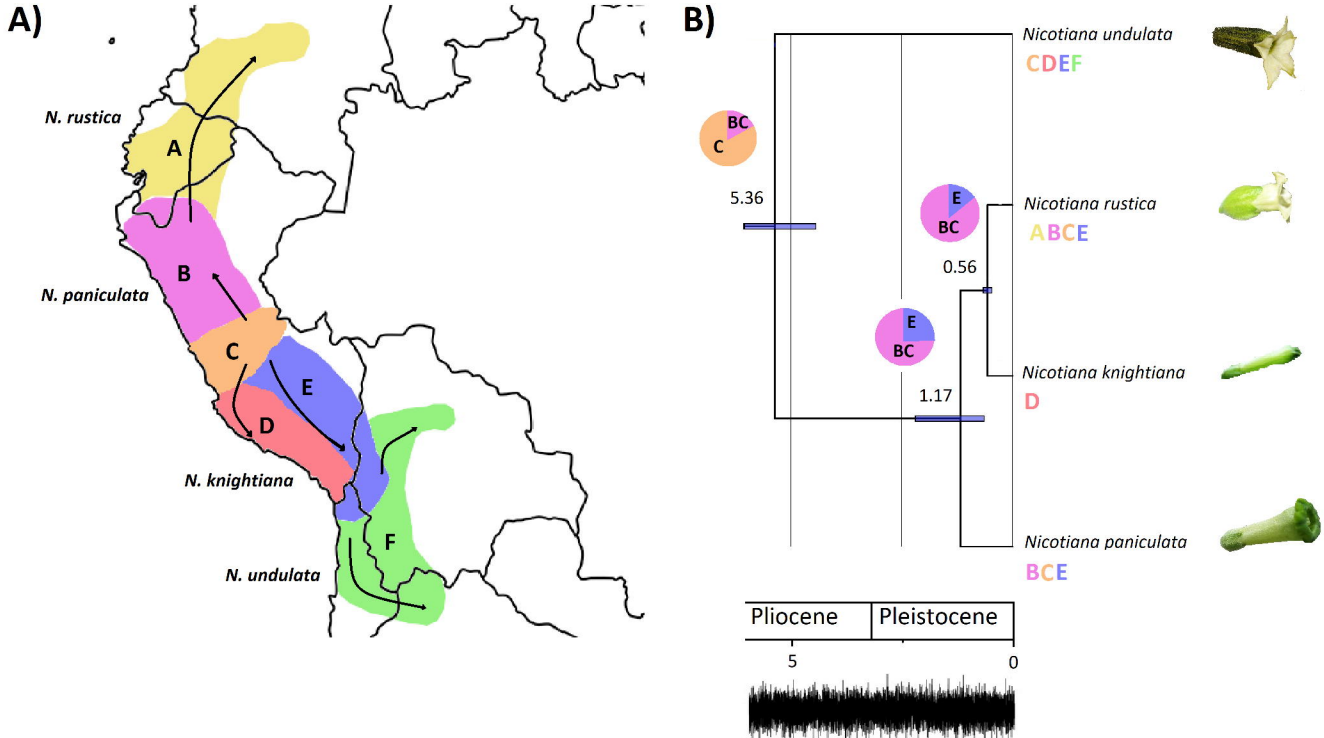

27.5
27
26.5
26
25.5
25
24.5
24
23.5
23
22.5
Number of tandem repeats
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum
Name of plant
A
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum


18
16
14
12
10
8
6
4
2
0
No. of repeats
IGS
CDS
Intron
Gene
Location of repeats
B
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Number of tandem repeats
20
22
26
28
30
32
34
36
38
40
42
44
48
50
54
63
64
72
84
88
Size of tandem repeats
C
Nicotiana knightiana
Nicotiana rustica
Nicotiana paniculata
Nicotiana obtusifolia
Nicotiana glauca
Nicotiana tabacum


Nucleotide Diversity
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
trnH-psbA
psbA
psbA-trnK-UUU
trnK-UUU-...
matK
matK-trnK-UUU-...
trnK-UUU-rps16
rps16-exon2
rps16-intron
rps16-trnQ-UUG
trnQ-UUG-psbK
psbK
psbK-psbI
psbI-trnS-GCU
trnS-GCU-trnG-...
trnG-UCC Intron
trnG-UCC-trnR-...
trnR-UCU-atpA
atpA
atpF Exon2
atpF intron
atpF exon1
atpF-atpH
atpH
atpH-atpI
atpI
atpI-rps2
rps2
rps2-rpoC2
rpoC2
rpoC1 exon2
rpoC1 intron
rpoC1 exon1
rpoB
rpoB-trnC-GCA
trnC-GCA-petN
petN-psbM
psbM-trnD-GUC
trnD-GUC-trnY-...
trnE-UUC-trnT-...
trnT-psbD
psbD
psbC
psbC-trnS-UGA
trnS-UGA-psbZ
psbZ
psbZ-trnG-GCC
trnG-GCC-trnfM
trnfM-rps14
rps14
rps14-psaB
psaB
psaA
psaA-ycf3
ycf3 exon3
ycf3 intron 2
ycfr3 exon 2
ycf3 intron 1
ycf3 exon1
ycf3-trnS-GGA
trnS-GGA-rps4
rps4
rps4-trnT-UGU
trnT-UGU-trnL-...
trnL-UAA-trnF-...
trnF-GAA-ndhJ
ndhJ
ndhJ-ndhK
ndhK
ndhC
ndhC-trnV-UAC
trnM-CAU-atpE
atpE
atpB
atpB-rbcL
rbcL
rbcL-accD
accD
accD-psaI
0.3
0.25
0.2
0.15
0.1
0.05
0
psaI-ycf4
ycf4
ycf4-cemA
cemA
cemA-petA
petA
petA-psbJ
psbJ
psbJ-psbL
psbL
psbF
psbE-petL
petL
petL-petG
petG-trnW-CCA
trnW-CCA-trnP-UGG
trnP-UGG-psaJ
psaJ
psaJ-rpl33
rpl33-rps18
rps18
rps18-rpl20
rpl20
rpl20-rps12
rps12 exon 2
rps12 intron
rps12 exon 1
clpP
clpP-psbB
psbB
psbB-psbT
psbT-psbN
psbH
psbH-petB
petB
PetB-petD
petD
petD-rpoA
rpoA
rpoA-rps11
rps11
rpl36-infA
infA
infA-rps8
rps8
rps8-rpl14
rpl14
rpl14-rpl16
rpl16
rps3
rpl22
rps19
rps19-rpl2
rpl2
ycf2
ycf15-trnL-CAU
ndhB
rps12-trnV-GAC
rrn4.5-rrn5
rrn5-trnR
trnR-ACG-trnN-GUU
trnN-GUU-ycf1
ndhF
ndhF-rpl32
rpl32-trnL-UAG
ccsA
ccsA-ndhD
ndhD
ndhD-psaC
psaC
psaC-ndhE
ndhE
ndhE-ndhG
ndhG
ndhG-ndhI
ndhI
ndhA
ndhH
ndhH-rps15
rps15
rps15-ycf1
ycf1


Solanum dulcamara
Nicotiana otophora
100
Nicotiana tomentosiformis
Nicotiana glauca
100
Nicotiana sylvestris
100
Nicotiana tabacum
90
Nicotiana obtusifolia
100
Nicotiana attenuata
100
Nicotiana undulata
100
Nicotiana paniculata
100
Nicotiana rustica
100
Nicotiana knightiana
0.004


A)
N. rustica
A
B
N. paniculata
C
D
E
N. knightiana
F
N. undulata
B)
Nicotiana undulata
CDEF
BC
C
5.36
Nicotiana rustica
ABCE
E
BC
0.56
E
BC
Nicotiana knightiana
D
1.17
Nicotiana paniculata
BCE
Pliocene
Pleistocene
5
0