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, 2, Abdullah Abdullah 1, Zartasha Ubaid 1, Iram Shahzadi 1, Ibrar Ahmed 3, Mohammad Tahir Waheed 1, Peter Poczai Corresp., 2, Bushra Mirza Corresp. 1

1 Department of Biochemistry, Quaid-i-Azam University, Islamabad, Pakistan
2 Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland
3 Alpha Genomics Private Limited, Islamabad, Pakistan

Corresponding Authors: Peter Poczai, Bushra Mirza
Email address: peter.poczai@helsinki.fi, bushramirza@qau.edu.pk

Species of the genus *Nicotiana* (Solanaceae), commonly referred to as tobacco plants, are often cultivated as non-food crops and garden ornamentals. In addition to the worldwide production of tobacco leaves, they are also used as evolutionary model systems due to their complex development history, tangled by polyploidy and hybridization. Here, we assembled the plastid genomes of five tobacco species: *N. knightiana*, *N. rustica*, *N. paniculata*, *N. obtusifolia* and *N. glauca*. *De novo* assembled tobacco plastid genomes had the typical quadripartite structure, consisting of a pair of inverted repeat (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 with currently available Solanaceae genome sequences showed similar GC content, gene content, codon usage, simple sequence repeats, oligonucleotide repeats, RNA editing sites, and substitutions. We identified 20 highly polymorphic regions, mostly belonging to intergenic spacer regions (IGS), which could be suitable for the development of robust and cost-effective markers for inferring the phylogeny of the 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*. Relaxed molecular clock analyses estimated the speciation event between *N. rustica* and *N. knightiana* appeared 0.56 Ma (HPD 0.65–0.46). Biogeographical analysis supported 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 indicate that four genes involved in different
plastid functions, including DNA replication (rpoA) and photosynthesis (atpB, ndhD and
ndhF), came under positive selective pressure as a result of specific environmental
conditions. Genetic mutations in these genes might have contributed to better survival and
superior adaptations during the evolutionary history of tobacco species.
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 Mehmood1,2, Abdullah1, Zartasha Ubaid1, Iram Shahzadi1, Ibrar Ahmed3, Mohammad Tahir Waheed1, Péter Poczai*3, Bushra Mirza*1

1Department of Biochemistry, Quaid-i-Azam University, Islamabad, Pakistan
2Finnish Museum of Natural History (Botany Unit), University of Helsinki, Helsinki, Finland
3Alpha Genomics Private Limited, Islamabad, Pakistan

*Corresponding authors: Bushra Mirza (bushramirza@qau.edu.pk)
Péter Poczai (peter.poczai@helsinki.fi)
Abstract

Species of the genus *Nicotiana* (Solanaceae), commonly referred to as tobacco plants, are often cultivated as non-food crops and garden ornamentals. In addition to the worldwide production of tobacco leaves, they are also used as evolutionary model systems due to their complex development history, tangled by polyploidy and hybridization. Here, we assembled the plastid genomes of five tobacco species: *N. knightiana*, *N. rustica*, *N. paniculata*, *N. obtusifolia* and *N. glauca*. *De novo* assembled tobacco plastid genomes had the typical quadripartite structure, consisting of a pair of inverted repeat (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 with currently available Solanaceae genome sequences showed similar GC content, gene content, codon usage, simple sequence repeats, oligonucleotide repeats, RNA editing sites, and substitutions. We identified 20 highly polymorphic regions, mostly belonging to intergenic spacer regions (IGS), which could be suitable for the development of robust and cost-effective markers for inferring the phylogeny of the 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*. Relaxed molecular clock analyses estimated the speciation event between *N. rustica* and *N. knightiana* appeared 0.56 Ma (HPD 0.65–0.46). Biogeographical analysis supported 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 indicate that four genes involved in different plastid functions, including DNA replication (*rpoA*) and photosynthesis (*atpB*, *ndhD* and *ndhF*), came under positive selective pressure as a result of specific environmental conditions. Genetic mutations in these genes might have contributed to better survival and superior adaptations during the evolutionary history of tobacco species.

Keywords: Mutational hotspots, *Nicotiana*, plastid genomes, positive selection, substitutions, speciation
1. Introduction

*Nicotiana* L. is the fifth largest genus in the megadiverse plant family Solanaceae, comprising 75 species (Olmstead et al., 2008; Olmstead & Bohs, 2007), 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 that *Nicotiana* consists of 13 sections (Knapp, Chase & Clarkson, 2004). One significant utilization of *Nicotiana* species has been as a source of genetic diversity for improving one of the most widely cultivated non-food crops, common tobacco (*N. tabacum* L.). This species is of major economic interest and is grown worldwide for its leaves used in the manufacture of cigars, cigarettes, pipe tobacco, and smokeless tobacco products consumed by more than one billion people globally (Lewis., 2011; Occhialini et al., 2016). While *N. tabacum* is the most notable commercial species for commerce, several additional species are also cultivated for smoking (*N. rustica* L.) and ornamental (*N. sylvestris* Spegazzini & Comes) or industrial (*N. glauca* Graham) purposes (Lester and Hawkes., 2001). Aztec or Indian tobacco (*N. rustica*), characterized by short yellowish flowers and round leaves, is widely cultivated in Mexico and North America. It was the first tobacco species introduced to Europe in the 16th century, but later superseded by *N. tabacum* for its milder taste (Shaw, 1960). Known as “o-yen’-kwa hon’we” (real tobacco) by North American Iroquois (Kell, 1966), it was used for medicinal and ritual purposes or even in weather forecasting (Winter, 2001). Aztec tobacco is still cultivated in South America, Turkey, Russia and Vietnam due to its tolerance to adverse climatic conditions (Sierro et al., 2018).

Some members of *Nicotiana* offer several research advantages, including extensive phenotypic diversity, amenability to controlled hybridizations and ploidy manipulations, high fecundity, and excellent response to tissue culture (Lewis, 2011). Consequently, *N. tabacum* and *N. benthamiana* Domin have become model organisms in the generation of new knowledge related to hybridization, cytogenetics, and polyploid evolution (Goodin et al., 2008; Zhang et al., 2011; Bally et al., 2018; Schiavinato et al., 2019). The first complete plastid genome sequence to be published for *N. tabacum* (Shinozaki et al., 1986). Since then, the structure and composition of plastid genomes has become widely utilized in identifying unique genetic changes and the evolutionary
relationships of various groups of plants. Furthermore, plastid genes have also been linked with important crop traits such as yield and resistance to pests and pathogens (Jin & Daniell, 2015).

Chloroplasts (cp) are large, double-membrane organelles with a genome size of 75-250 kb (Palmer, 1985). Most chloroplast encoded proteins are responsible for photosynthesis and for the synthesis of fatty acids and amino acids (Cooper, 2000). Angiosperm plastid genomes commonly contain ~130 genes, comprised of 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- plastid genomes typically quadripartite structure, consisting of 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., 2019a). Numerous mutational events occur in plastid genomes, including variations in tandem repeats, insertions and deletions (indels), and point mutations; inversions and translocations are also common (Jheng et al., 2012; Xu et al., 2015; Abdullah et al., 2020).

The plastid genome of angiosperms has maternal inheritance (Daniell, 2007), which together with its conserved organization makes it extremely useful for exploring phylogenetic relationships at different taxonomic levels (Ravi et al., 2008). Plastid genome polymorphisms are useful for species barcoding, solving taxonomic issues, studying population genetics, and for investigating species adaptation to their natural habitats (Ahmed, 2014; Daniell et al., 2016; Nguyen et al., 2017). 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 negative or purifying selection to maintain their function, while positive selection might act on some genes in response to environmental changes (Iram et al., 2019; Henriquez et al., 2020a).

Nicotiana species are diploid (2n = 2x = 24), although allopolyploid species are also common in the genus (Leitch et al., 2008). Phylogenetic studies have shown that these allopolyploids were formed 0.4 million (N. rustica and N. tabacum) (Clarkson, Dodsworth & Chase, 2017) to 5 million years ago (species of sect. Suaveolentes) (Schiavinato et al., 2020). Nicotiana tabacum (2n=4x=48), is known to be a natural allopolyploid derived from two closely related ancestors (Lim et al., 2007). The paternal donor N. tomentosiformis L. (2n=24) was confirmed by genomic
in situ hybridization (GISH) (Clarkson et al., 2005), physical mapping (Bindler et al., 2011) and
genome sequencing (Sierro et al., 2014), while the maternal donor N. sylvestris (2n=24) was
identified by plastid genome sequencing (Yukawa, Tsudzuki & Sugiura, 2006). Aztec tobacco (N.
rustica), like N. tabacum, is also an allotetraploid but has originated from the recent hybridization
of different parental species. Based on morphology, karyotype analyses and crossing experiments,
Goodspeed (1954) suggested the ancestral species are N. paniculata L. (2n=2x=24; maternal) and
N. undulata Ruiz & Pav. (2n=2x=24; paternal). The identity of the parental species was
investigated using nuclear internal transcribed spacer (ITS) regions, chloroplast markers in situ
hybridization methods, and genome sequencing (Aoki & Ito, 2000; Chese et al., 2003; Clarkson et
al., 2004; Lim et al., 2004, 2007; Kovarik et al., 2004; Sierro et al., 2018). These analyses
confirmed N. undulata as the paternal ancestor according to Goodspeed’s hypothesis and showed
the genome of N. rustica lacks inter-genomic translocations (Lim et al., 2004, Kovarik et al., 2012),
while additivity can be observed in the 5S and 35S rDNA loci respect to its progenitors (Lim et
al., 2007), which were homogenized by concerted evolution (Kovarik et al., 2004). These analyses
did not provide further evidence for the maternal parent of N. rustica but suggested either N.
knightiana L. or N. paniculata L. could be the maternal donor.

Here, we assembled the plastid genome of five Nicotiana species and compared their sequences to
gain insight into the plastid genome structure of genus Nicotiana. We also inferred the
phylogenetic relationship of the genus and investigated the selection pressures acting on protein-
coding genes. We 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. Lastly,
we used this information to trace the origin of the maternal genome of the allopolyplloid Nicotiana
rustica.

2. Materials and Methods
2.1. Plastid genome assembly and annotation
Illumina sequence data of N. knightiana (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 plastid genome sequence reads were selected by performing BWA-MEM
mapping with default settings (Li & Durbin, 2009) using Nicotiana tabacum (GenBank accession
number: NC_001879) as a reference. Geneious R8.1 (Kearse et al., 2012) de novo assembler was used to order the selected contigs for final assembly by selecting option “Medium sensitivity/Fast”, while keeping other parameters as default. Gene annotation was conducted using GeSeq (Tillich et al., 2017) with a BLAT (Kent, 2002) search of 85% to annotate protein-coding genes, rRNAs and tRNAs; CPGAVAS2 was used with default parameters by selecting option 1 “43-plastome” (Shi et al., 2019). After automatic annotation, start/stop codons and the position of introns were further confirmed manually by visual inspection of the translated protein of each gene in Geneious R8.1 and BLAST search using default settings with homologous genes of plastid genomes of family Solanaceae. The tRNA genes were further verified by tRNAscan-SE version 2.0 with default settings using options: sequence source “Organellar tRNA”, search mode “Default”, genetic code “Universal”, and Cut-off score for reporting tRNAs “15” (Lowe & Chan, 2016); ARAGORN version 1.2.38 was used with default parameters by selecting genetic code “Bacterial/Plant chloroplast” with maximum intron length of 3,000-bp (Laslett & Canback, 2004). Circular genome maps were drawn with OGDRAW v1.3.1 (Greiner, Lehwark & Bock, 2019) by uploading the GenBank (.gb) format of each plastid genome and selecting options: “Circular”, “Plastid”, “Tidy up annotation”, and “Draw GC graph”. The average coverage depth of Nicotiana species plastid genomes was calculated by mapping all raw reads without trimming to de novo assembled plastid genomes with BWA-MEM (Li & Durbin, 2009) visualized in Tablet (Milne et al., 2009). Novel Nicotiana plastid genomes were deposited in NCBI (Table S1).

2.2. Comparative genome analysis and RNA editing prediction

Novel plastid genome sequences were compared through multiple alignments using MAFFT v7 (Katoh & Standley, 2013). All parts of the genome, including intergenic spacer regions (IGS), introns, protein-coding genes, and ribosomal RNAs and tRNAs, were 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 compared to the N. tabacum reference using Geneious R8.1. 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. Structural borders of plastid genomes were compared for 10 selected Nicotiana species using IRscope with option “GB file upload” and default settings (Amiryousefi, Hyvönen & 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 codon usage and amino acid frequencies were determined by Geneious R8.1.

2.3. Repeats analyses

Microsatellites within the plastid genomes of five *Nicotiana* species were identified using MISA (Beier et al., 2017) with a minimal repeat number of 7 for mononucleotide repeats, 4 for dinucleotide repeats and 3 for tri-, tetra-, penta- and hexanucleotide SSRs. We also used REPuter software (Kurtz et al., 2001) with minimal repeat size set to 30 bp, Hamming distance set to 3, minimum similarity percentage of two repeat copies up to 90%, and a maximum computed repeat of 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. Non-synonymous (*K*$_{a}$) and synonymous (*K*$_{s}$) substitution rate analysis

To determined *K*$_{a}$ and *K*$_{s}$, protein-coding genes were extracted from the newly assembled *Nicotiana* plastid genomes and aligned using MAFFT with the corresponding genes of the previously published plastid genome of *Nicotiana tabacum* (Z00044.2) as a reference and analyzed using DnaSP v6. The data were interpreted in terms of purifying selection (*K*$_{a}$/*K*$_{s}$ < 1), neutral evolution (*K*$_{a}$/*K*$_{s}$ = 1), and positive selection (*K*$_{a}$/*K*$_{s}$ > 1).

We evaluated 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. From all the analyses performed in DATAMONKEY, the most suited model of evolution for each dataset was selected as directly estimated on this web server. 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 with variable $\omega$ across lineages at individual sites, restricting $\omega$ to $\leq 1$ 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 p values...
< 0.05 using the false discovery rate (FDR) correction according to Benjamini & Hochberg (1995) in Microsoft Excel.

### 2.5. Phylogenetic analyses

Plastid genome sequences of the genus *Nicotiana* were selected from the Organelle Genome Resources of the NCBI (accessed on 21.02.2019) and used in phylogenetic inference along with de novo assembled sequences of *Nicotiana*. The x=12 clade includes the traditional subfamily Solanoideae plus *Nicotiana* with the Australian endemic Anthocercideae tribe and takes its name from the synapomorphy of chromosome numbers based on 12 pairs (Olmstead et al., 2008). *Nicotiana* and Anthocercideae appear to be in a first branching position in the x=12 clade, thus we have chosen *Solanum dulcamara* L. (Amiryousefi, Hyvönen & Poczai, 2018a) from the Solanoideae tribe with a curated plastid genome as an outgroup for rooting our phylogenetic tree.

For the species included in our analysis, coding alignments were constructed from the excised plastid genes using MACSE (Ranwez et al., 2011), including 75 protein-coding genes (Table S2). For phylogenetic analysis, a 75,449-bp concatenated matrix was used with the best fitting GY+F+I+G4 model determined by ModelFinder (Kalyaanamoorthy et al., 2017) 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. The key idea behind UFBoot is to keep trees encountered during the ML-tree search for the original sequence alignment and to use them to evaluate the tree likelihoods for the bootstrap sequence alignment. UFBoot provides relatively unbiased bootstrap estimates under mild model misspecifications and reduces computing time while achieving more unbiased branch supports than with standard bootstrap (Hoang et al., 2018). TreeDyn was used for further enhancement of the phylogenetic tree analysis (Dereeper et al., 2008; Lemoine et al., 2019).

Relative divergence times were estimated for 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 an uncorrelated relaxed 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). 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 with Markov Chain Monte Carlo (MCMC) samples based on 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 the Köppen-Geiger climate classification 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); and (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, 2020) 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

The genome size of the assembled complete plastid genomes ranged between 155,689 bp (N. paniculata) and 156,022 bp (N. obtusifolia), while reads provided 327 to 1,951x coverage (Table S1). Genomes harbored 133 unique genes, of which 18 genes were duplicated in the IR region (Table S2, Fig.1a). Out of 133 genes, 85 were protein-coding, 37 were tRNA genes and 8 were
rRNA genes. Among 18 duplicated genes in the IR region, 7 were protein-coding, 7 were tRNA genes and 4 were rRNA genes. From the protein-coding genes, 18 contained introns, while rps12 was a trans-spliced gene with its 1st exon found in the LSC and the 2nd and 3rd exons found in the IR region. Structural elements of the IR region also showed the highest GC content (43.2%) compared to the LSC (35.9%) and SSC (32.1%) (Table S1). This finding could be attributed to the presence of tRNA (52.9%) and rRNA (55.4%) genes in inverted repeats.

The nucleotide composition comparison of Nicotiana genomes revealed high synteny among all regions, including not only the LSC, SSC, IR and CDSs, but interestingly also in non-coding regions. Detailed comparison of the base composition of each region is shown in Table S3. All amino acid sequences in Nicotiana plastid genomes were rich in AT bases and coded a higher percentage of hydrophobic amino acids compared to acidic amino acids that (Fig. 1b). Codon usage and frequency of amino acids revealed that leucine is the most abundant and cysteine the least encoded amino acid in these genomes (Fig. 1b). At the 3rd codon position the frequency of A/T codons was higher compared to C/G (Table S4).

The number of RNA editing sites predicted using PREP-cp varied between 34 and 37, distributed among 15 genes (see Table S3). From these genes, the most RNA-edited sites were possessed by ndhB (9), 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, with the frequency of second base codon edits being much higher. These changes helped in the formation of hydrophobic amino acids, for example valine (V), leucine (L) and phenylalanine (F), with conversions from serine to leucine being the most frequent. (Table S5).

3.2. IR contraction and expansion

The JL (LSC/IR) and JS (IR/SSC) border positions of Nicotiana plastid genomes were compared (Fig. 1c) using IRscope (Amiryousefi, Hyvönen & Poczai, 2018b). The length of the IR regions was similar, ranging from 25,331 bp to 25,436 bp, with some expansion. The JLA (IRa/LSC) junction point was located between trnH-GUG and rpl2 among Nicotiana plastid genomes. In N. tomentosiformis and N. attenuata, the IR expanded to partially include rps19, creating a 60 and 54-bp truncated pseudogenic rps19 copy at JLA (IRa/LSC). Furthermore, infA, ycf15, and a copy
of ycf1 located on the JSB (IRb/SSC) were detected as pseudogenes. The position of ycf1 in the IRb/SSC region varied. It left a 33-bp pseudogene in *N. obtusifolia*, a 36-bp pseudogene in *N. knightiana*, *N. rustica* and *N. glauca* and a 72-bp pseudogene in *N. paniculata*.

### 3.3. Non-synonymous (*K_a*) and synonymous (*K_s*) substitution rate analysis

The synonymous/non-synonymous substitution rate 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 *N. tabacum* as a reference. 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, 13 genes showed *K_a*/*K_s* > 1 in at least one species (Table S4). We selected these genes 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 ω = 1, ω < 1 or ω > 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 *atpB*, *ndhD*, *ndhF* and *rpoA* (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).

### 3.4. 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 the species. The majority of the SSRs in these plastid genomes were mononucleotide rather than trinucleotide or dinucleotide repeats. The most dominant of the SSRs were mononucleotide A/T motifs, while the second most predominant were dinucleotide AT/TA motifs. Mononucleotide SSRs varied from 7- to 17-unit repeats, dinucleotide SSRs varied from 4- to 5-unit repeats, and other SSRs types present mainly in 3-unit repeats. Most SSRs were located in the LSC and were less frequent in the IR and SSC (Fig. S1 and Table S7). Locating repeats with REPuter revealed 117 oligonucleotide repeats evenly dispersed among the
species, ranging from 21 (N. paniculata) to 25 (N. knightiana and N. glauca). Forward (F) and palindromic (P) repeats were abundant in all species, where N. glauca had the lowest number of repeats [9 (39%) (F) and 11 (52%) (P)] and N. obtusifolia harbored the highest number of repeats [14 (56%) (F) and 11 (44%) (P)]. The size of oligonucleotide repeats varied from 30–65 bp, and many were found in the intergenomic spacer regions (IGS) of the LSC (Fig. S2 and Table S8). The non-coding IGS regions also contained most of the tandem repeats (Fig. S3).

3.5. Single nucleotide polymorphism and insertion/deletion analyses in Nicotiana

To discover polymorphic regions (mutational hotspots), the CDS, intron and IGS regions of the whole plastid genome of five Nicotiana species were compared. Nucleotide diversity values varied from 0 (ycf3) to 0.306 (rps12 intron region) (Fig. 2). High polymorphism was found in intronic regions (average \( \pi = 0.1670 \)) compared to IGS (\( \pi = 0.031 \)) and CDS regions (average \( \pi = 0.002 \)). We further investigated the number and occurrence of substitution types in the plastid genomes using N. tabacum as a reference and encountered 509 (N. galuca) to 861 (N. paniculata) substitutions along the entire plastid genome. Most of the conversions were A/G and C/T single nucleotide polymorphisms (SNPs) (Table 2). A detailed description of the ratio of transition to transversion substitutions (Ts/Tv) can be found in Table S9. In addition to the distribution of SNPs, we examined insertions and deletions (indels) and located 107 (N. rustica) to 143 (N. obtusifolia) polymorphisms across the compared genomes (Table 3). Based on this comparison we successfully determined 20 highly polymorphic regions that might be used as potential markers in Nicotiana species barcoding (Table 4).

3.6. Phylogenetic analyses

Phylogenetic analysis within Nicotiana plastid genomes was performed with the maximum likelihood method based on 75 selected and concatenated protein-coding genes. Our phylogenetic analyses resulted in a highly resolved tree (Fig. 1d). Almost all the recovered clades have maximum branch support values reconstructed based on alignment size of 75,449 bp with best fit model GY+F+I+G4. 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 million years ago (Ma) (highest posterior density, HPD 6.38–4.43), while N. paniculata diverged 1.17 Ma (HPD 2.18–0.63) followed by the most recent split of N. rustica and N.
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. 3). 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 plastid genomes from five *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 novel *Nicotiana* plastid genomes were similar to those previously reported (Sugiyama et al., 2005; Yukawa, Tsudzuki & Sugiura, 2006); the GC content was high in the IR, which might be a result of the presence of ribosomal RNA (Qian et al., 2013; Cheng et al., 2017; Zhao et al., 2018). The genome organization, gene order and content also showed high similarity and synteny with sequences previously published for *N. slyvestris* and *N. tabacum* (Sugiyama et al., 2005; Yukawa, Tsudzuki & Sugiura, 2006). This could be attributed to plastid genomes of land plants having a conserved structure but with diversity prevailing at the border position of LSC/SSC/IR. However, examining the IR junction sites of *Nicotiana* species also showed similarities with some variation prevalent in *N. tomentosiformis* species, which has 60 bp in the IRb region, while the *rps19* gene is present entirely in the LSC region compared to the *N. tabacum* reference. Such fluctuations at the border positions of various regions of the plastid genome might be helpful in determining evolutionary relationships or could be indicators of environmental adaptation of species (Menezes et al., 2018). Liu et al., (2018) reported that the similarities in the junction regions may be useful for explaining the relationship between species, and that plants with a high level of relatedness show minimal fluctuations in the junctions of the plastid genome. In this respect, the high resemblance of the IR junction sites reveals a close relationship of *Nicotiana* species.
Repeats in the plastid genome are useful in evolutionary studies and play a vital role in genome arrangement (Zhang et al., 2016). We detected the presence of large amounts of mononucleotide repeats (A/T), and trinucleotide SSRs (ATT/TAA) in the five analyzed species of Nicotiana, which may be a result of the A/T-richness of the plastid genome. A similar result was also reported in N. 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 plastid genomes (Asaf et al., 2016; Shahzadi et al., 2019; Mehmood et al., 2019, Yang et al., 2019). To understand molecular evolution, it is important to analyze nucleotide substitution rates (Muse & Gaut, 1994); in plastid genomes LSC and SSC regions are more prone to substitutions and indels, whereas the IRs are more conserved (Ahmed et al., 2012; Abdullah et al., 2019b). Our results corroborate this finding, indicating the IR region is more conserved, and most of the substitutions occur in the LSC and SSC regions. Similar results have been shown in the plastid genome of yam (Dioscorea polystachya Turcz.) (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 evolution studies (Ahmed et al., 2013; Ahmed, 2014; Nguyen et al., 2017). Previous studies identified several polymorphic loci based on a comparison of plastid genomes, which have provided suitable information for the development of further molecular markers (Choi, Chung & Park, 2016; Li et al., 2018; Menezes et al., 2018). We found 20 polymorphism loci in Nicotiana that were more polymorphic than the frequently used rbcL, and matK markers. For example, infA, rps12 intron and rps16-trnQ-UUG had nucleotide diversities of 0.2594, 0.1527 and 0.0845, respectively. These regions show great potential as markers for population genetics and phylogenetic analyses 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 different environmental stresses. Nicotiana, originating in South America, has spread to many regions of the world and members of the genus have successfully adapted to survive in harsh environmental conditions. This large variation in their distributional range has induced distinctive habits and morphology in 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 pleiochasia
cymes to solitary flowers and diffuse panulate-cymose mixtures. For example, members of *Nicotiana*
section *Suaveolentes*, 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). *Nicotiana* is thus a rich
reservoir of genetic resources for evolutionary biological research since several members of the
genus have gone through changing climatic events and adapted 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, 2006).
$K_a/K_s$ values $>1$, $=1$, and $<1$ indicate positive selection, neutral evolution and purifying selection,
respectively (Lawrie et al., 2013). Many proteins and RNA molecules encoded by plastid genomes
have undergone 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 the case of plastid genes might
happen in response to certain environmental changes when advantageous genetic mutations can
contribute to better survival and adaptation. The $K_a/K_s$ ratios in our analysis indicate changes in
selective pressures. The genes *atpB*, *ndhD*, *ndhF* and *rpoA* had greater $K_a/K_s$ values (> 1), possibly
due to positive selective pressure as a result of specific environmental conditions. This was
conclusively supported by an integrative analysis using Fast Unconstrained Bayesian
AppRoximation (FUBAR) and Mixed Effects Model of Evolution (MEME) methods, which
identified a set of positively selected codons in these genes.

These genes are involved in different plastid functions, such as DNA replication (*rpoA*) and
photosynthesis (*atpB*, *ndhD* and *ndhF*). The *rpoA* 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 *rpoA* polycistron is essential for
chloroplast gene expression and plant development (Zhang et al., 2018). The housekeeping gene
\textit{atpB} encodes the \(\beta\)-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 \textit{ndh} genes encoding 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
different 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 demand for ATP is higher than for NADPH (e.g., during
photosynthetic induction, at high or low temperatures, 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 \textit{Nicotiana}.

These findings are further supported by our observation that RNA editing sites occur frequently in
\textit{Nicotiana ndh} genes (Table S3). It has been shown that \textit{ndhB} mutants under lower air humidity
conditions or following exposure to ABA present reduced levels of photosynthesis, 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 \textit{Nicotiana} species experienced abiotic stresses during their evolution,
which resulted in positive selection of some plastid genes. Up to this point, positive selection has
rarely been detected in plastid genes except for \textit{clpP} (Erixon & Oxelman, 2008), \textit{ndhF} (Peng et al.,
2011), \textit{matK} (Hao, Chen & Xiao, 2010) and \textit{rbcL} (Kapralov et al., 2011). However, Piot et al.
(2018) showed that one-third of the plastid genes in 113 species of grasses (Poaceae) evolved under positive selection. This indicates that positive selection is overlooked among diverse groups of plant taxa.

4.3. Phylogenetic relationships and the origin of tetraploid \textit{Nicotiana rustica}

Our comparative plastid genome analysis revealed that the maternal parent of the tetraploid \textit{N. rustica} was the common ancestor of \textit{N. paniculata} and \textit{N. knightiana}, with the latter species being more closely related to \textit{N. rustica}. The relaxed molecular clock analyses estimated that the speciation event between \textit{N. rustica} and \textit{N. knightiana} appeared \textasciitilde{}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 \textit{Nicotiana} species revealed that \textit{N. rustica} inherited about 41\% of its nuclear genome from its paternal progenitor, \textit{N. undulata}, and the rest from its maternal progenitor, the common ancestor of \textit{N. paniculata} and \textit{N. knightiana} (Sierro et al., 2018), which has also been confirmed by our study. We also revealed that \textit{N. knightiana} is more closely related to \textit{N. rustica} than \textit{N. paniculata}, which can be further corroborated by the distribution of indels highlighted in the present study.

The paternal inheritance of plastid genomes was observed in \textit{Nicotiana} under certain stressed conditions (Medgyesy, Fejes & Maliga, 1985; Medgyesy, Páy & Márton, 1986; Thang and Medgyesy, 1989; Avni and Edelman, 1991; Ruf, Karcher & Bock, 2007; Thyssen, Svab & Maliga, 2012). Such low-frequency paternal leakage of plastids via pollen was suggested to be universal in plants with strict maternal plastid inheritance (Azagiri and Maliga 2007). Thus, we expect that the plastids in the putative parents of \textit{N. rustica} are maternally inherited. Medgyesy, Páy & Márton (1986) observed the paternal transmission of plastids in \textit{N. plumbaginifolia} Viv., but they concluded that plants carried maternal mitochondria. Further studies investigating the parental origins of \textit{Nicotiana} species should also focus on mitochondrial genomes excluding possible low-frequency paternal plastid inheritance.

The biogeographical analysis suggests that \textit{N. undulata} and \textit{N. paniculata} evolved in North/Central Peru, while \textit{N. rustica} developed in Southern Peru and separated from \textit{N. knightiana}, which adapted to the Southern coastal climatic regimes. Positively selected plastid genes with functions such as DNA replication (\textit{rpoA}) and photosynthesis (\textit{atpB}, \textit{ndhD} and \textit{ndhF}) might have been associated with successful adaptation to, for example, a coastal environment. However, our results are tentative, as our study lacks data for several broad ecological variables, including variation in
salinity, island versus mainland, and East versus West of the Andes. We aim to highlight that many potential environmental variables might be highly correlated with speciation processes in *Nicotiana*, as has been demonstrated in the same region for another Solanaceae group in the tomato clade (*Solanum sect. Lycopersicon*), where amino acid differences in genes associated with seasonal climate variation and intensity of photosynthetically active radiation have been correlated with speciation processes (Pease et al., 2016). Another example of rapid adaptive radiation from the family is the genus *Nolana* L.f., where several clades gained competitive advantages in water-dependent environments by succeeding and diverging in Peru and Northern Chile (Dillon et al., 2009). In the case of *N. rustica* and related species, we assume that diversification was driven by the ecologically variable environments of the Andes. Our molecular clock analysis provides evidence for recent species diversification in the Pleistocene and Pliocene while substantial climatic transitions in Peru predate these events. For example, the uplift of the central region of the Andes and the formation of the Peruvian coastal desert ended ~14 – 150 Mya (Hoorn et al., 2010; Gerreaud, Molina & Farias, 2010), before the geographical and ecological expansion of *N. rustica* and related parental species.

The dispersal of *N. rustica* and related species shows a south-to-north range expansion and diversification which has been suggested by phylogenies of other plant and animal groups in the Central Andes (Picard, Sempere & Plantard, 2008; Lueber & Weigend, 2014). Based on the south-to-north progression scenario, habitats located at high altitudes were first available for colonization in the south, recently continuing to northward. Erosion and orogenic progression caused dispersal barriers of species colonizing these high habitats to diversify in a south-to-north pattern, frequently following allopatric speciation. Thus, for taxonomic groups currently residing throughout a large portion of the high Andes, a south-to-north speciation pattern is expected (Doan, 2003). In this case, the most basal species (*N. undulata*) has more southern geographic ranges, and the most derived species (*N. rustica*) has more northern geographic ranges, except for *N. knightiana*, which presumably colonized the coastal range of Peru. Although the four *Nicotiana* species examined show overlaps in their distribution, it is probable that speciation was caused by fragmentation of populations during the glacial period (see Simpson, 1975). Utilizing fewer chloroplast loci for phylogenetic analyses of plant species may limit the solution of 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 plastid genomes has provided deep insight into the phylogeny of certain families and genera (Henriquez et al., 2014; Amiryousefi, Hyvönen & Poczai, 2018a; Abdullah et al., 2020). Here, we reconstructed a phylogenetic tree for 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 *N. tabacum* of section *Nicotiana*, and *N. sylvestris* of section *Sylvestres*, show close resemblance. *N. paniculata* and *N. knightiana* belong to section *Paniculatae* and appeared to reflect the maternal ancestry of these species relative to *N. rustica* of section *Rusticae*. Overall, our phylogenetic analyses support the previous classification of genus *Nicotiana*, and corroborates that plastid genomic resources can provide further support for highly resolved phylogenies.

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 plastid genome was similar to those of previously reported Solanaceae plastid genomes. 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 and Kathryn Rannikko for editing the manuscript.

Competing interest

The authors declare that they have no conflict of interest.

Authors contributions

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

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

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

Ibrar Ahmed: Conceptualization, Manuscript editing.

Mohammad Tahir Waheed: Conceptualization, Manuscript editing.

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

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

Péter Poczai and Bushra Mirza, contributed equally to the manuscript.

References

Abdullah, Mehmood F, Shahzadi I, Waseem S, Mirza B, Ahmed I, Waheed MT. 2020. Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): Comparative analyses and identification of mutational hotspots. *Genomics* **112**:581–591. DOI: 10.1016/j.ygeno.2019.04.010.

Abdullah, Shahzadi I, Mehmood F, Ali Z, Malik MS, Waseem S, Mirza B, Ahmed I, Waheed MT. 2019a. Comparative analyses of chloroplast genomes among three *Firmiana* species: Identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae. *Plant Gene* **19**:100199

Abdullah, Waseem S, Mirza B, Ahmed I, Waheed MT. 2019b. Comparative analyses of chloroplast genome in *Theobroma cacao* and *Theobroma grandiflorum*. *Biologia* DOI: https://doi.org/10.2478/s11756-019-00388-8

Ahmed I. 2014. *Evolutionary dynamics in taro*. Massey University, Palmerston North, New Zealand, PhD Thesis
Ahmed I, Biggs PJ, Matthews PJ, Collins LJ, Hendy MD, Lockhart PJ. 2012. Mutational dynamics of aroid chloroplast genomes. *Genome Biology and Evolution* **4**:1316–1323. DOI: 10.1093/gbe/evs110.

Ahmed I, Matthews PJ, Biggs PJ, Naeem M, Mclenachan PA, Lockhart PJ. 2013. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. *Molecular 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.

Bally J, Jung H, Mortimer C, Naim F, Philips JG, Hellens R, Bombarely A, Goodin MM, Waterhouse PM. 2018. The rise and rise of *Nicotiana benthamiana*: a plant for all reasons. *Annual Review of Phytopathology* **56**: 405-426. DOI: 10.1146/annurev-phyto-080417-050141.

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.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**:289–300.
Bindler G, Plieske J, Bakaher N, Gunduz I, Ivanov N, van der Hoeven R, Ganal M, Donini P. 2011. A high density genetic map of tobacco (Nicotiana tabacum L.) obtained from large scale microsatellite marker development. Theoretic and Applied Genetics 123: 219

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.

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.

Clarkson JJ, Lim KY, Kovarik A, Chase MW, Knapp S, Andrew RL. 2005. Long-term genome diploidization in allopolyploid Nicotiana section Repandae (Solanaceae). New Phytologist 168:241–252

Clarkson JJ, Dodsworth S, Chase MW. 2017. Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in Nicotiana (Solanaceae). Plant Systematics and Evolution 303:1001–1012. DOI: 10.1007/s00606-017-1416-9.

Clarkson JJ, Knapp S, Garcia VF, Olmstead RG, Leitch AR, Chase MW. 2004. Phylogenetic relationships in Nicotiana (Solanaceae) inferred from multiple plastid DNA regions. Molecular Phylogenetics and Evolution 33:75–90.
Cooper G. 2000. *Chloroplasts and other plastids in the cell: A molecular approach*. Sunderland (MA): Sinauer Associates.

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 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 *Proctopus* (Reptilia, Gymnophthalmidae). *Journal of Biogeography* 30: 361–374

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

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

Erixon P, Oxelman B. 2008. Whole-gene positive selection, elevated synonymous substitution
rates, duplication, and indel evolution of the chloroplast clpP1 gene. *PLoS ONE* 3: e1386. DOI: 10.1371/journal.pone.0001386.

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

Gatenby AA, Rothstein SJ, Nomura M. 1989. Translational coupling of the maize chloroplast *atpB* and *atpE* genes. *Proceedings of the National Academy of Sciences USA* 86: 4066–4070.

Gerreaud RD, Molina A, Farias M. 2010. Andean uplift, ocean cooling and Atacama hyperaridity: a climate modeling perspective. *Earth and Planetary Science Letters* 292: 39–50.

Goodin MM, Zaitlin D, Naidu RA, Lommel SA. 2008. *Nicotiana benthamiana*: Its history and future as a model for plant-pathogen interactions. *Molecular Plant-Microbe Interactions* 21: 1015-1026. DOI: 10.1094/MPMI-21-8-1015.

Goodspeed TH. 1954. *The Genus Nicotiana*. Chronica Botanica, New York, USA.

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

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

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

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.

Henriquez CL, Abdullah, Ahmed I, Carlsen MM, Zuluaga A, Croat TB, Mckain MR. 2020a. Molecular evolution of chloroplast genomes in Monsteroideae (Araceae). *Planta* 251: 72.
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, Sanmartín I, Sanchez-Meseguer A, Anderson CL, Figueiredo JP, Jaramillo C, Riff D, Negri FR, Hooghiemstra H, Lundberg J, Stadler T, Särkinen T, Antonelli A. 2010. Amazonia through time: 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.

Iram S, Hayat MQ, Tahir M, Gul A, Abdullah, Ahmed I. 2019. Chloroplast genome sequence of *Artemisia scoparia*: Comparative analyses and screening of mutational hotspots. *Plants* 8:476. DOI: doi:10.3390/plants8110476.

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. The Engineered Chloroplast Genome Just Got Smarter. *Trends in Plant Science* 20: 622–640 DOI: 10.1016/j.tiplants.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.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780. DOI: 10.1093/molbev/mst010.

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

Kent WJ. 2002. BLAT - The BLAST-like alignment tool. Genome Research 12: 656–664. DOI: 10.1101/gr.229202. Article published online before March 2002.

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

Kimura M. 2006. Model of effectively neutral mutations in which selective constraint is incorporated. Proceedings of the National Academy of Sciences USA 76: 3440–3444. DOI: 10.1073/pnas.76.7.3440.

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

Kurtz S, Choudhuri J V, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research 29:4633–4642.

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

Kell KT. 1966. Folk Names for Tobacco. The Journal of American Folklore 79:590–599. DOI: 10.2307/538224.
Kovarik A, Matyasek R, Lim KY, Skalická K, Koukalová B, Knapp S, Chase M, Leitch AR. 2004. Concerted evolution of 18-5.8-26S rDNA repeats in Nicotiana allotetraploids. Biological Journal of the Linnean Society 82:615–625 DOI: 10.1111/j.1095-8312.2004.00345.x.

Kovarik A., Renny-Byfield S., Grandbastien MA., Leitch A. 2012. Evolutionary Implications of Genome and Karyotype Restructuring in Nicotiana tabacum L. In: Soltis P., Soltis D. (eds) Polyploidy and Genome Evolution. Springer, Berlin, Heidelberg pp.209–224

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

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in D. melanogaster. PLoS Genetics 9:33–40. DOI: 10.1371/journal.pgen.1003527.

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

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

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

Li Y, Zhang Z, Yang J, Lv G. 2018. Complete chloroplast genome of seven Fritillaria species, variable DNA markers identification and phylogenetic relationships within the genus. PLoS ONE 13:e0194613 DOI: 10.1371/journal.pone.0194613. 3.

Lim KY, Matyasek R, Kovarik A, Leitch AR. 2004. Genome evolution in allotetraploid Nicotiana. Biological Journal of the Linnean Society 82:599–606 DOI: 10.1111/j.1095-8312.2004.00344.x. 82:599–606.

Lim KY, Kovarik A, Matyasek R, Chase MW, Clarkson JJ, Grandbastien MA, Leitch AR. 2007. Sequence of events leading to near-complete genome turnover in allopolyploid
Nicotiana within five million years. New Phytologist 175:756–763. DOI: 10.1111/j.1469-8137.2007.02121.x.

**Liu L, Wang Y, He P, Li P, Lee J, Soltis DE, Fu C. 2018.** Chloroplast genome analyses and genomic resource development for epilithic sister genera Oresitrophe and Mukdenia (Saxifragaceae), using genome skimming data. BMC Genomics 19:235. DOI: 10.1186/s12864-018-4633-x.

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

**Lewis RS. 2011.** Nicotiana. In: Kole C (ed) Wild crop relatives: genomic and breeding resources plantation and ornamental crops. Springer, New York, USA.

**Lester RN, Hawkes JG. 2001.** Solanaceae. In: Hanelt P and Institute of Plant Genetics and Crop Plant Research (eds) Mansfeld’s encyclopedia of agriculture and horticultural crops (Except Oranamentals), vol 4. Springer, Berlin. pp. 1790–1856.

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

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

**Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoeb B, Hasegawa M, Penny D. 2002.** Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proceedings of the National Academy of Sciences of the USA 99:12246–51. DOI: 10.1073/pnas.1824329999.

**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 112: 1522–1530 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, 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

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..

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

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

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.

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:
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. Molecular phylogeny of the Solanaceae. *Molecular Phylogenetics and Evolution* 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* : 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.

Peng L, Yamamoto H, Shikanai T. 2011. Structure and biogenesis of the chloroplast NAD (P)H dehydrogenase complex. *Biochimica et Biophysica Acta* 1807: 946–953

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
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]

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

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

Särkinen T, Bohs L, Olmstead RG, Knapp S. 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae) a dated 1000-tip tree. BMC Evolutionary Biology 13:214

Schiavinato M, Marcet-Houben M, Dohm JC, Gabaldón T, Himmelbauer H. 2020. Parental origin of the allotetraploid tobacco Nicotiana benthamiana. Plant Journal DOI: 10.1111/tpj.14648.

Schiavinato M, Strasser R, Mach L, Dohm JC, Himmelbauer H. 2019. Genome and
transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔxT/FT.

*BMC Genomics* 20: 594 DOI: 10.1186/s12864-019-5960-2

Shahzadi I, Abdullah, Mehmood F, Ali Z, Ahmed I, Mirza B. 2019. Chloroplast genome sequences of *Artemisia maritima* and *Artemisia absinthium*: Comparative analyses, mutational hotspots in genus *Artemisia* and phylogeny in family Asteraceae. *Genomics* 112:1454–1463 DOI: 10.1016/J.YGENO.2019.08.016

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

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

Sierro N, Battey JND, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV (2014) The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications* 5:3833

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

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

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M. 2005. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: Comparative analysis of mitochondrial genomes in higher plants. *Molecular Genetics and Genomics* 272:603–615 DOI: 10.1007/s00438-004-1075-8.

Shaw T. 1960. Early Smoking Pipes: In Africa, Europe, and America. *The Journal of the Royal
Terashima I, Araya T, Miyazawa SI, Sone K, Yano S. 2005. Construction and maintenance of the optimal photosynthetic systems of the leaf, herbaceous plant and tree: An eco-developmental treatise. *Annals of Botany* 95:507–519 DOI: 10.1093/aob/mci049.

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

Winter JC. 2001. Traditional uses of tobacco by native Americans. In: Winter JC (ed) Tobacco use by native North Americans, University of Oklahoma Press, Norman, USA. pp. 57.

Turrill WB, Goodspeed TH. 2007. The Genus *Nicotiana*. *Kew Bulletin* 10: 357–358 DOI: 10.2307/4109222.

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

Xu S, Bröckmöller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, Zhou W, Kreitzer C, Stanke M, Tang H, Lyons E, Pandey P, Pandey SP, Timmermann B, Gaquerel E, Baldwin IT. 2017. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proceedings of the National Academy of Sciences USA* 114:6133–6138

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

Yamori W, Shikanai T. 2016. Physiological functions of cyclic electron transport around photosystem I in sustaining photosynthesis and plant growth. *Annual Review of Plant 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. 2020. RASP 4: ancestral state reconstruction tool for multiple genes and characters. *Molecular Biology and Evolution* 37:604–606.
Yukawa M, Tsuzuki 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 *rpoA* polycistron for plastid-encoded RNA polymerase-dependent chloroplast gene expression and chloroplast development. *Scientific Reports* 8:1–12 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* 7:1–12. 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. (a) Chloroplast genome map of five *Nicotiana* species. Genes lying outside the circle are transcribed clockwise; genes inside the circle are transcribed counterclockwise. Different colors indicate functional groups the genes belong to. GC genome content is plotted light gray and AT dark gray 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. Genes with introns are marked with asterisks. (b) Comparison of amino acid groups and amino acid frequencies in six *Nicotiana* species. (c) Comparison of the border positions of LSC, SSC and IR among five *Nicotiana* chloroplast genomes. Positive strand transcribed genes are indicated below the line; negative strand transcribed genes are indicated above the line. Gene names are contained in boxes, and the lengths of relative regions are above the boxes. The number of bp (base pairs) that are written with genes reveal the part of the gene that exists in the region of the chloroplast or away from the region of the chloroplast i.e. bp written with *ycf1* indicates that sequences exist in that region of the plastid genome. (d) Maximum likelihood (ML) tree reconstructed from 75 protein-coding plastid genes of 11 *Nicotiana* species with *Solanum dulcamara* as an outgroup. Bootstrap support values are shown above or below the nodes.

Figure 2. Nucleotide diversity of different regions of the chloroplast genome among *Nicotiana* species. The x-axis indicates the chloroplast regions; y-axis indicates nucleotide diversity.

Figure 3. 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 *Nicotiana rustica* in South America. Arrows illustrate dispersal events inferred from the biogeographic analysis. Geographical distribution for each terminal is indicated using the biogeographic region’s 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 displayed above each node. All nodes have PP ≥ 0.97 and BS ≥ 87%. Trace plot of the combined chains shows 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.

Table 2. Comparison of substitutions in *Nicotiana* species.

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

Table 4. Mutational hotspots among *Nicotiana* species.
Figure 1

Chloroplast genome map of five *Nicotiana* species.

(a) Chloroplast genome map of five *Nicotiana* species. 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. Genes with introns are marked with asterisks. (b) Comparison of amino acid groups and amino acid frequencies in six *Nicotiana* species. (c) 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. (d) 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 2

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 3

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 ≥ 0.97 and BS ≥ 87%. Trace plot of the combined chains showing the sampled joint probability and the convergence of the chains.
Table 1 (on next page)

Summary statistics

Summary statistics of de novo assembled Nicotiana plastid genomes
**Table 1.** List of amino acid replacements and results of positive selection tests on codons underlying these replacements.

| Gene | Position | α     | β     | Amino acid replacements | FUBAR (PP) | MEME (LRT) | FDR  |
|------|----------|-------|-------|-------------------------|------------|------------|------|
| atpB | 19       | 0.909 | 6.339 | K N K K N N N N N N     | 0.918      | 4.42       | 0.044|
|      | 21       | 0.697 | 15.511| P P P P P P P P P P   | 0.990      | 4.99       | 0.012|
| ndhD | 153      | 1.981 | 12.245| C C C C C C C C C C C | 0.927      | 3.16       | 0.021|
|      | 185      | 0.744 | 9.365 | V L V V V V L L V V V | 0.956      | 3.82       | 0.017|
| ndhF | 460      | 1.385 | 8.772 | V V V V V V V A A V V  | 0.912      | 3.30       | 0.010|
| rpoA | 201      | 1.149 | 8.426 | L L L L L L L L L L   | 0.940      | 3.84       | 0.023|

Note: α – the mean synonymous substitution rate at a site; β – the mean non-synonymous substitution rate at a site; PP – posterior probability of positive selection at a site; LRT – Likelihood ratio test for episodic (positive) diversification; FDR – false discovery rate (FDR=5%).

Species abbreviations: Nat - *N. attenuata*; Ngla - *N. glauca*; Nkni - *N. knightiana*; Nobt - *N. obtusifolia*; Noto - *N. otophora*; Npan - *N. paniculata*; Nrus - *N. rustica*; Nsyl - *N. sylvestris*; Ntab - *N. tabacum*; Ntom - *N. tomentosiformis*; Nund - *N. undulata*.
Table 2 (on next page)

Comparison of substitution in *Nicotiana* species

Comparison of substitution in *Nicotiana* species. *Nicotiana tabacum* was used as reference for SNPs detection.
1 Table 2. Comparison of substitution in *Nicotiana* species

| Types | *Nicotiana knightiana* | *Nicotiana rustica* | *Nicotiana paniculata* | *Nicotiana obtusifolia* | *Nicotiana glauca* |
|-------|------------------------|---------------------|------------------------|------------------------|-------------------|
| A/G   | 222                    | 219                 | 245                    | 244                    | 110               |
| C/T   | 226                    | 223                 | 237                    | 250                    | 128               |
| A/C   | 105                    | 104                 | 117                    | 153                    | 97                |
| C/G   | 40                     | 39                  | 50                     | 52                     | 34                |
| G/T   | 130                    | 129                 | 135                    | 148                    | 87                |
| A/T   | 63                     | 61                  | 77                     | 102                    | 53                |
| Total | 786                    | 775                 | 861                    | 847                    | 509               |

**Location wise distribution**

|       | LSC | SSC | IR  |
|-------|-----|-----|-----|
| A/G   | 560 | 559 | 43  |
| C/T   | 630 | 671 | 32  |
| A/C   | 183 | 184 | 33  |
| C/G   | 198 | 210 | 68  |
| G/T   | 32  | 82  |
| A/T   | 100 |

2 *Nicotiana tabacum* was used as reference for SNPs detection.
**Table 3** (on next page)

InDels distribution

InDels distribution of *Nicotiana* chloroplast genome
### Table 3. Distribution of InDels in *Nicotiana* chloroplast genome

| Species             | InDel length (bp) | InDel average length |
|---------------------|-------------------|----------------------|
|                     | LSC               | SSC                  | IR       |
| *Nicotiana knightiana* | 91                | 11                   | 8        |
|                     | 506               | 36                   | 29       |
|                     | 5.56              | 3.27                 | 3.62     |
| *Nicotiana rustica*  | 89                | 11                   | 7        |
|                     | 478               | 36                   | 38       |
|                     | 5.37              | 3.27                 | 5.42     |
| *Nicotiana paniculata* | 92               | 14                   | 10       |
|                     | 618               | 156                  | 28       |
|                     | 6.71              | 11.14                | 2.80     |
| *Nicotiana obtusifolia*  | 117              | 12                   | 14       |
|                     | 677               | 52                   | 167      |
|                     | 5.78              | 4.33                 | 11.92    |
| *Nicotiana glauca*   | 88                | 11                   | 14       |
|                     | 450               | 44                   | 82       |
|                     | 5.11              | 4                    | 5.85     |
## Table 4 (on next page)

**Mutational hotspots**

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

| S. No | Region           | Nucleotide Diversity | T. No's of mutation | Region Length |
|-------|------------------|----------------------|---------------------|---------------|
| 1     | *infA*           | 0.25942              | 45                  | 249           |
| 2     | rps12 intron     | 0.15275              | 161                 | 527           |
| 3     | rps16-trnQ-UUG   | 0.08451              | 225                 | 1266          |
| 4     | trnK-UUU-rps16   | 0.0483               | 46                  | 703           |
| 5     | trnH-psbA        | 0.04387              | 19                  | 433           |
| 6     | rpl36-infA       | 0.02942              | 3                   | 116           |
| 7     | ccsA-ndhD        | 0.02876              | 17                  | 237           |
| 8     | rps16-intron     | 0.02784              | 27                  | 862           |
| 9     | rpl32-trnL-UAG   | 0.02617              | 61                  | 931           |
| 10    | trnM-CAU-atpE    | 0.02244              | 24                  | 218           |
| 11    | rpl33-rps18      | 0.02222              | 20                  | 180           |
| 12    | petD-rpoA        | 0.01989              | 9                   | 182           |
| 13    | rpl14-rpl16      | 0.01849              | 10                  | 119           |
| 14    | ndhE-ndhG        | 0.01735              | 7                   | 219           |
| 15    | rps15-ycfI       | 0.01714              | 17                  | 385           |
| 16    | ndhH-rps15       | 0.01667              | 4                   | 108           |
| 17    | petG-trnW-CCA    | 0.01575              | 4                   | 127           |
| 18    | petL-petG        | 0.01538              | 6                   | 182           |
| 19    | trnG-GCC-trnfM   | 0.01524              | 11                  | 228           |
| 20    | rpoA-rps11       | 0.01515              | 2                   | 66            |