




OPEN

DATA DESCRIPTOR

# Chromosome-level genome assemblies of *Nicotiana tabacum*, *Nicotiana sylvestris*, and *Nicotiana tomentosiformis*

Nicolas Sierro , Mehdi Auberson, Rémi Dulize & Nikolai V. Ivanov


The Solanaceae species *Nicotiana tabacum*, an economically important crop plant cultivated worldwide, is an allotetraploid species that appeared about 200,000 years ago as the result of the hybridization of diploid ancestors of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. The previously published genome assemblies for these three species relied primarily on short-reads, and the obtained pseudochromosomes only partially covered the genomes. In this study, we generated annotated *de novo* chromosome-level genomes of *N. tabacum*, *N. sylvestris*, and *N. tomentosiformis*, which contain 3.99 Gb, 2.32 Gb, and 1.74 Gb, respectively of sequence data, with 97.6%, 99.5%, and 95.9% aligned in chromosomes, and represent 99.2%, 98.3%, and 98.5% of the near-universal single-copy orthologs Solanaceae genes. The completion levels of these chromosome-level genomes for *N. tabacum*, *N. sylvestris*, and *N. tomentosiformis* are comparable to other reference Solanaceae genomes, enabling more efficient synteny-based cross-species research.

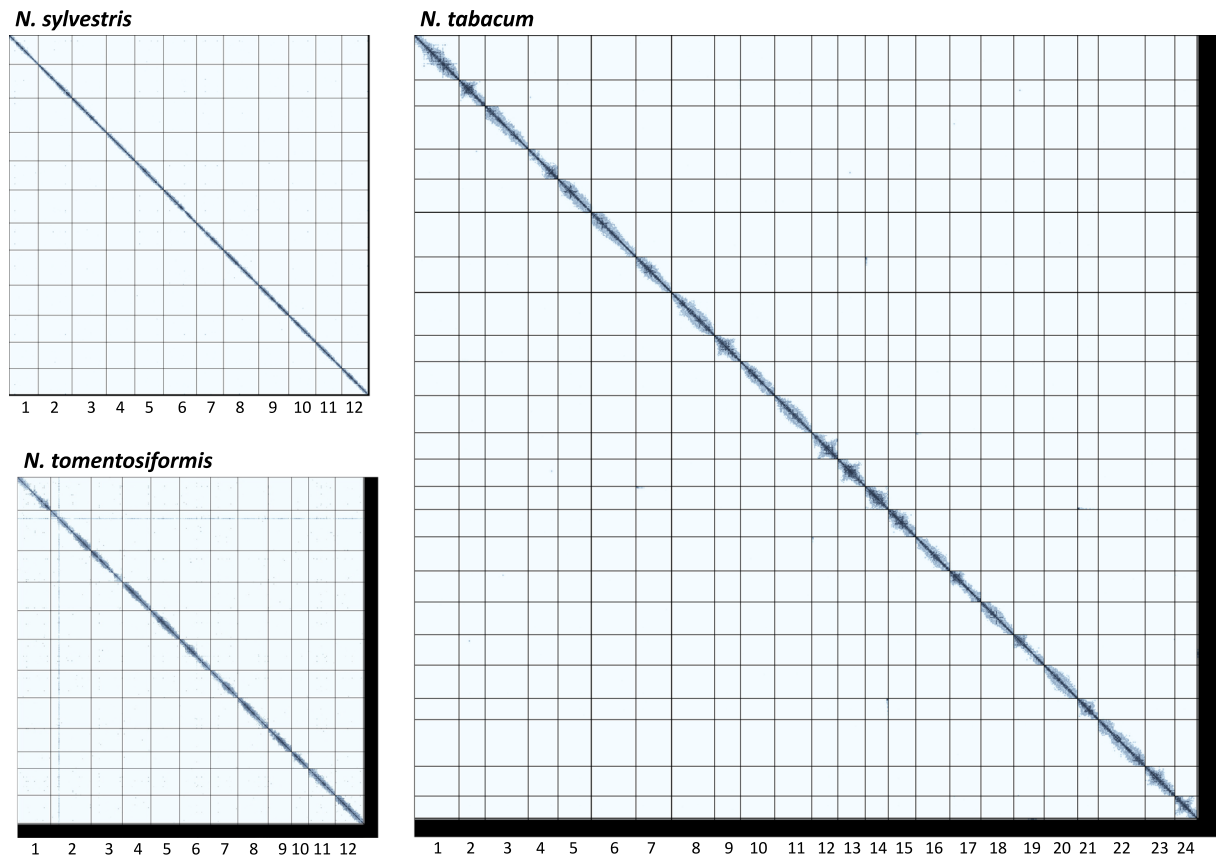
## Background & Summary

The *Nicotiana* genus belongs to the Solanaceae family, which also includes tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), and eggplant (*Solanum melongena*)<sup>1,2</sup>. While most of the Solanaceae are diploids with 12 chromosome pairs, tobacco (*Nicotiana tabacum* L.) is an allotetraploid ( $2n = 4x = 48$ ) resulting from a hybridization event that likely occurred in the Andes within the last 200,000 years between ancestors of *Nicotiana sylvestris* (S-genome;  $2n = 2x = 24$ ) and *Nicotiana tomentosiformis* (T-genome;  $2n = 2x = 24$ )<sup>3,4</sup>. In addition to being a modern descendant of the *N. tabacum* maternal progenitor, *N. sylvestris*, which is nowadays largely cultivated as an ornamental plant, is also one of the closest descendants of the ancestral species from the *Alatae/Sylvestres* section that hybridized as the paternal donor with an ancestral species from the *Noctiflorae/Petunioides* section to give rise to the almost all-Australian clade of allopolyploid species constituting the *Nicotiana* section *Suaveolentes*<sup>5</sup>.

Similar to other members of the *Nicotiana* genus, *N. sylvestris*, *N. tomentosiformis*, and *N. tabacum* produce a wide range of alkaloids that are known to be toxic to insects and are a well-established mechanism of defense against herbivores<sup>6</sup>. While *N. sylvestris* accumulates similar amounts of alkaloids in roots and leaves (3.5 mg/g in roots and 2.1 mg/g in leaves), *N. tomentosiformis* accumulates more alkaloids in roots (8.8 mg/g in roots and 0.6 mg/g in leaves), and *N. tabacum* has more in leaves (1.3 mg/g in roots and 12.5 mg/g in leaves)<sup>7</sup>. The composition of the accumulated alkaloids varies between the three species, with *N. tabacum* benefiting from both of its progenitors' genetic and regulatory contributions. In *N. sylvestris* roots, 87% of the alkaloids is nicotine, 11% is anatabine, and 1.9% is anabasine, while in leaves, 100% of the alkaloids is nicotine. In *N. tomentosiformis* roots, 56% of the alkaloids is nornicotine, 28% is anatabine, 14% is nicotine, 1.6% is anabasine, and 0.57% is cotinine, while in leaf 73% of the alkaloids is nicotine and 27% is nornicotine. In *N. tabacum* roots, 87% of the alkaloids is nicotine, and 13% is nornicotine, while in leaves, 92% of the alkaloids is nicotine, 5.1% is nornicotine, and 2.6% is anatabine<sup>7</sup>.

The *Nicotiana* genus is also a rich source of terpenoids, which play a significant role as attractants to several pollinator insects. In *N. tabacum*, both cembranoid and labdanoid diterpenoids are synthesized in the trichome

PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, CH-2000, Neuchâtel, Switzerland.  e-mail: [nicolas.sierro@cgat.ch](mailto:nicolas.sierro@cgat.ch)



**Fig. 1** PoreC contact maps. Intra-chromosomal and inter-chromosomal contacts are shown for the *Nicotiana sylvestris*, *Nicotiana tomentosiformis*, and *Nicotiana tabacum* genome assemblies. The black bottom and right edges correspond to unplaced sequences.

glands, whereas *N. sylvestris* produces predominantly cembranoid diterpenoids and *N. tomentosiformis* predominantly labdanoid diterpenoids<sup>8</sup>.

Although several *Nicotiana* species genomes have been published in the last decade, including for *N. sylvestris*<sup>9</sup>, *N. tomentosiformis*<sup>9</sup>, and *N. tabacum*<sup>10,11</sup>, these genomes are primarily based on the assembly of second-generation sequencing data and therefore suffer from an important fragmentation resulting in only partial anchoring to chromosomes.

In the present study, we integrated Illumina short-read sequencing (Illumina, San Diego, CA, USA) with third-generation Oxford Nanopore long-read sequencing and Oxford Nanopore chromosome conformation capture (PoreC) technology (Oxford Nanopore Technologies, Oxford, UK) to generate high-quality chromosome-level reference genomes for *N. tabacum*, *N. sylvestris*, and *N. tomentosiformis*. These new resources will broaden our understanding of the contributions of both *N. tabacum* progenitors to the genes and the pathways of tobacco and enable more efficient synteny-based cross-species Solanaceae research.

## Methods

**DNA Extraction and Sequencing.** Young leaves from *N. tabacum* L. Cultivar K326 (PVY resistant derived from USDA ARS GRIN Global NPGS: PI 552505), *N. Sylvestris* Speg. TW136 (USDA ARS GRIN Global NPGS: PI 555569) and *N. tomentosiformis* Goodsp. TW142 (USDA ARS GRIN Global NPGS: PI 555572) were snap-frozen with liquid nitrogen and finely ground in a mortar. High molecular weight genomic DNA for long-read sequencing was extracted using Promega Wizard HMW DNA Extraction Kit (Promega AG, Madison, WI, USA).

Short genomic DNA fragments were deleted using Circulomics short-read eliminator kits from PacBio (PacBio, Menlo Park, CA, USA), and long-read sequencing libraries were prepared using Oxford Nanopore Technologies SQK-LSK109 Ligation Sequencing Kits before sequencing on Oxford Nanopore Technologies PromethION R9.4.1 flowcells. About 139 Gb of raw data were collected for *N. tabacum*, 159 Gb for *N. sylvestris*, and 76 Gb for *N. tomentosiformis*.

To conduct chromosome-level assembly, frozen leaves were cut into one square centimeter pieces and treated with formaldehyde to fix the DNA. The fixed genomic DNA was then digested overnight using the NlaIII restriction enzyme, and the 3' overhangs were re-ligated using T4 ligase before extraction. PoreC sequencing libraries were prepared using Oxford Nanopore Technologies SQK-LSK109 Ligation Sequencing Kits before sequencing on Oxford Nanopore Technologies PromethION R9.4.1 flowcells. About 40 Gb of raw data were collected for *N. tabacum*, 66 Gb for *N. sylvestris*, and 63 Gb for *N. tomentosiformis*.

	<i>N. sylvestris</i>	<i>N. tomentosiformis</i>	<i>N. tabacum</i>
Chr01	188,594,255	159,904,673	222,086,288
Chr02	216,772,750	195,009,794	130,336,781
Chr03	222,355,857	151,524,687	215,112,738
Chr04	182,174,535	137,929,616	150,061,924
Chr05	183,858,385	139,001,852	166,564,982
Chr06	213,680,027	150,012,402	218,894,441
Chr07	174,301,312	131,654,490	180,540,203
Chr08	226,073,828	146,876,285	212,334,375
Chr09	193,471,688	113,607,955	128,373,858
Chr10	173,459,395	80,117,918	173,455,358
Chr11	166,895,819	128,153,117	182,075,898
Chr12	168,333,862	136,938,251	131,632,239
Chr13			135,455,135
Chr14			117,149,023
Chr15			135,283,689
Chr16			171,654,204
Chr17			153,689,555
Chr18			161,231,186
Chr19			151,880,626
Chr20			168,699,142
Chr21			101,618,745
Chr22			234,076,610
Chr23			146,657,309
Chr24			112,906,552
Unplaced	11,613,273	71,022,666	93,985,334
Total	2,321,584,986	1,741,753,706	3,995,756,195
% anchored	99.5%	95.9%	97.6%

**Table 1.** Chromosome length, total assembly length, and percentage of the assembly anchored to chromosomes for *Nicotiana sylvestris*, *Nicotiana tomentosiformis*, and *Nicotiana tabacum*.

To polish and validate the assembled genomes, Illumina short-reads were prepared for *N. tabacum* using Tecan Celero EZ DNA-Seq Library Preparation Kits (Tecan, Männedorf, Switzerland) and sequenced as 2 × 151 bp paired-end reads on an Illumina NovaSeq 6000 to generate a total of 139 Gb. Illumina short-reads from ERR274527<sup>12</sup> and ERR274528<sup>13</sup> for *N. sylvestris* and from ERR274540<sup>14</sup> and ERR274542<sup>15</sup> for *N. tomentosiformis* were retrieved from the Short Read Archive.

**De novo Assembly and Chromosome Construction.** For *N. tabacum*, Oxford Nanopore basecalling was performed using Guppy 6.3.7 using the plant super model. Long-read sequences were filtered using seqkit<sup>16</sup> 2.2.0 to remove short (length <5000) and low-quality reads (average qscore <9), resulting in 98 Gb (N50 length: 28.5 kb).

For *N. sylvestris* and *N. tomentosiformis*, Oxford Nanopore basecalling was performed using Guppy 6.1.1 using the plant super model. Long-read sequences were filtered using seqkit<sup>16</sup> 2.2.0 to remove short (length <2500) and low-quality reads (average qscore <9), resulting in 108 Gb (N50 length: 25.9 kb) and 41 Gb (N50 length: 28.2 kb) for *N. sylvestris* and *N. tomentosiformis*, respectively.

Genomes were assembled using flye<sup>17</sup> 2.9.1 using the nano-hq input pre-set and a read error rate of 0.03.

The Illumina short-reads were processed for each species using fastp<sup>18</sup> 0.23.2 to trim adapters and low-quality bases, merge pairs, and remove low complexity and short (length <75) reads. During processing, the reads were split into two sets, one for assembly polishing which contained 80% of the processed Illumina reads and one for assembly validation containing 20% of the processed Illumina reads.

The assembled genomes were polished with processed Illumina short-reads using fmlrc2<sup>19</sup> 0.1.7. The remaining haplotig sequences were removed from the assemblies using purge\_dups<sup>20</sup> 1.2.6, with cut-offs set to 3, 8, and 1000 for *N. tabacum*, to 5, 10, and 1000 for *N. sylvestris*, and to 2, 3, and 1000 for *N. tomentosiformis*.

Illumina short-reads were mapped to the assembly contigs using minimap2<sup>21,22</sup> 2.24, duplicates marked with samblaster<sup>23</sup> 0.1.26, and filtered using samtools<sup>24</sup> 1.15.1. The coverage of the assembly contigs by Illumina sequencing was then calculated using samtools<sup>24</sup> 1.15.1, and contigs with less than 70% of their length with a coverage of at least 5 for *N. tabacum* and 15 for *N. sylvestris* and *N. tomentosiformis* were removed.

Because the biological material used for sequencing originated from inbred plants that can be considered homozygotes, variants were called using freebayes<sup>25</sup> 1.3.6 with the ploidy parameter set to 1 and ignoring sites with coverage higher than 200 and filtered with vcfliib<sup>26</sup> 1.0.3 vcfliib using the parameters --filter-sites-info --filter "QUAL >20 & QUAL/AO >10 & SAF >0 & SAR >0 & RPL >1 & RPR >1". Variants were then applied to the genomes using bcftools<sup>24</sup> 1.15.1 consensus to generate the polished assembly contigs.

				<i>N. sylvestris</i>			<i>N. tomentosiformis</i>			<i>N. tabacum</i>		
				length	% of TE	% of genome	length	% of TE	% of genome	length	% of TE	% of genome
Class I	LINE			9,675,761	1.6%	0.4%	9,615,294	1.7%	0.6%	19,745,491	1.7%	0.5%
	LTR	Ty1/copia	Ale	13,558,929	2.2%	0.6%	11,670,567	2.1%	0.7%	24,896,186	2.1%	0.6%
	LTR	Ty1/copia	Alesia	299,556	0.0%	0.0%	102,813	0.0%	0.0%	380,375	0.0%	0.0%
	LTR	Ty1/copia	Angela	3,989,230	0.6%	0.2%	1,986,947	0.4%	0.1%	5,826,451	0.5%	0.1%
	LTR	Ty1/copia	Bianca	14,202,928	2.3%	0.6%	12,796,742	2.3%	0.7%	26,200,095	2.2%	0.7%
	LTR	Ty1/copia	Ikeros	2,115,094	0.3%	0.1%	1,471,616	0.3%	0.1%	3,637,043	0.3%	0.1%
	LTR	Ty1/copia	Ivana	1,366,613	0.2%	0.1%	1,775,454	0.3%	0.1%	3,081,268	0.3%	0.1%
	LTR	Ty1/copia	SIRE	24,828,773	4.0%	1.1%	14,684,980	2.6%	0.8%	38,903,674	3.3%	1.0%
	LTR	Ty1/copia	TAR	9,154,238	1.5%	0.4%	12,202,057	2.2%	0.7%	20,574,287	1.8%	0.5%
	LTR	Ty1/copia	Tork	14,900,248	2.4%	0.6%	7,980,408	1.4%	0.5%	21,895,949	1.9%	0.5%
	LTR	Ty3/gypsy	chromovirus CRM	2,617,599	0.4%	0.1%	2,797,010	0.5%	0.2%	5,522,890	0.5%	0.1%
	LTR	Ty3/gypsy	chromovirus Chlamyvir	0	0.0%	0.0%	0	0.0%	0.0%	5,786	0.0%	0.0%
	LTR	Ty3/gypsy	chromovirus Galadriel	5,700,997	0.9%	0.2%	5,500,007	1.0%	0.3%	11,145,772	1.0%	0.3%
	LTR	Ty3/gypsy	chromovirus Reina	2,340,197	0.4%	0.1%	2,730,019	0.5%	0.2%	4,966,891	0.4%	0.1%
	LTR	Ty3/gypsy	chromovirus Tcn1	0	0.0%	0.0%	0	0.0%	0.0%	3,227	0.0%	0.0%
	LTR	Ty3/gypsy	chromovirus Tekay	248,756,499	40.3%	10.7%	310,558,487	55.4%	17.8%	559,106,371	47.7%	14.0%
	LTR	Ty3/gypsy	chromovirus chromo-outgroup	0	0.0%	0.0%	5,356	0.0%	0.0%	15,594	0.0%	0.0%
	LTR	Ty3/gypsy	non-chromovirus OTA Athila	59,196,881	9.6%	2.5%	50,111,064	8.9%	2.9%	108,359,430	9.3%	2.7%
	LTR	Ty3/gypsy	non-chromovirus OTA Tat Ogre	116,167,517	18.8%	5.0%	21,672,795	3.9%	1.2%	135,653,424	11.6%	3.4%
	LTR	Ty3/gypsy	non-chromovirus OTA Tat Retand	75,130,400	12.2%	3.2%	81,189,167	14.5%	4.7%	155,002,488	13.2%	3.9%
	pararetrovirus			6,158,720	1.0%	0.3%	3,683,969	0.7%	0.2%	9,724,069	0.8%	0.2%
Class II	Subclass 1	TIR	EnSpm/CACTA	758,007	0.1%	0.0%	1,412,566	0.3%	0.1%	2,030,473	0.2%	0.1%
	Subclass 1	TIR	MuDR/Mutator	2,475,225	0.4%	0.1%	1,382,863	0.2%	0.1%	4,014,994	0.3%	0.1%
	Subclass 1	TIR	PIF/Harbinger	114,448	0.0%	0.0%	128,968	0.0%	0.0%	294,983	0.0%	0.0%
	Subclass 1	TIR	Tc1/Mariner	26,044	0.0%	0.0%	100,923	0.0%	0.0%	83,809	0.0%	0.0%
	Subclass 1	TIR	hAT	3,476,633	0.6%	0.1%	3,931,259	0.7%	0.2%	7,522,070	0.6%	0.2%
	Subclass 2	Helitron	Helitron	860,377	0.1%	0.0%	1,440,401	0.3%	0.1%	23,79,749	0.2%	0.1%
Total				617,870,914	100.0%	26.6%	560,931,732	100.0%	32.2%	1,170,972,839	100.0%	29.3%

**Table 2.** Predicted retrotransposons length and genome coverage statistics.

Assembly contigs from plastid and mitochondrion were removed by mapping the polished assembly contigs to the *N. tabacum* plastid and mitochondrion sequences (NC\_001879.2<sup>27</sup> and NC\_006581.1<sup>28</sup>, respectively) using minimap2<sup>21,22</sup> 2.24 and filtering out contig mapping on more than 50% of their length.

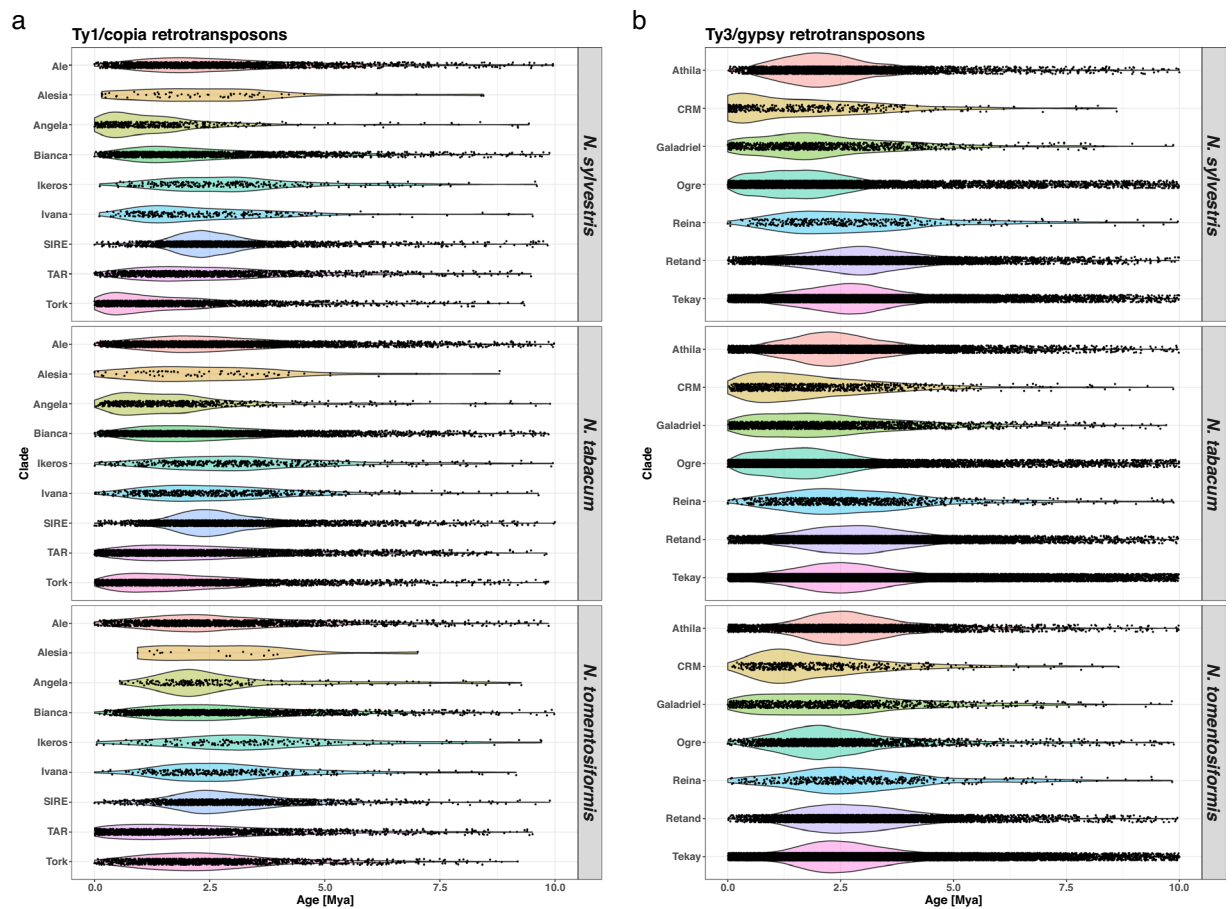
Assembly contigs from possible contamination were identified using kraken2<sup>29</sup> 2.1.2 using the k2\_plus\_pfp\_20220908 database<sup>30</sup> and removed by only retaining contigs identified as belonging to *Nicotiana* or *Solanum* species.

PoreC reads were mapped to the cleaned assembly contigs using minimap2<sup>21,22</sup> 2.24. Alignments with a mapping quality lower than 60 for *N. tabacum* and 30 for *N. sylvestris* and *N. tomentosiformis* were discarded, and contact pairs were created from the remaining alignments. The positions on the contigs of each contact pair were recorded as two consecutive lines in a BED file. The scaffolding of the contigs to a chromosome-level assembly was performed using yahs<sup>31</sup> 1.2a1. Contact maps were prepared using PretextMap<sup>32</sup> 0.1.9, manually curated and annotated in PretextView<sup>33</sup> 0.2.5, and the resulting scaffolds exported as chromosome-level sequences.

To name and orient the *N. tabacum* chromosome-level sequences, the PT markers, mapped to the sequences using hisat2<sup>34</sup> 2.2.1 and the tobacco genetic map<sup>35</sup>, were used. Similarly, the *N. tomentosiformis* chromosome-level sequences were named and oriented using the N genetic map<sup>36</sup> combined with the tobacco PT markers<sup>35</sup>. The chromosome-level assembly of the *N. tomentosiformis* genome was then used as a reference to name and orient the *N. sylvestris* chromosome-level sequences based on minimap2<sup>21,22</sup> 2.24 mapping (Fig. 1).

The proportion of the assembly anchored to chromosomes reached 99.5%, 95.9%, and 97.6% of the total assembly lengths for *N. sylvestris*, *N. tomentosiformis*, and *N. tabacum*, respectively (Table 1).

When compared to the previously available *N. tabacum* genome assembly<sup>11</sup> generated from short-read sequencing, whole genome profiling and optical and genetic mapping data, the new *N. tabacum* genome assembly has fewer contigs (decrease from 1,257,801 to 1410) with a larger N50 length (increase from 9.1 kb to 11.8 Mb), and the proportion of the assembly anchored to chromosomes consequently improved from 64% to 97.6%.



**Fig. 2** Predicted retrotransposon insertion ages. **(a)** Predicted insertion ages in millions of years for retrotransposons of the Ty1/copia superfamily; **(b)** Predicted insertion ages in millions of years for retrotransposons of the Ty3/gypsy superfamily.

**Retrotransposon Prediction and Annotation.** Nested retrotransposons were annotated by iteratively running genomertools 1.6.2 ltrharvest<sup>37</sup> using the parameters -similar 70 -seed 20 -minlenltr 100 -maxlenltr 7000 -mindistltr 1000 -maxdistltr 15000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 3 -vic 10 -overlaps best, retaining the predictions matching to the RepeatExplorer Viridiplantae 3.0 dataset<sup>38</sup> using diamond<sup>39</sup> 2.1.6 blastx with the parameters --max-target-seqs 1 --ultra-sensitive --frameshift 15, and excising them from the assembly using samtools<sup>24</sup> 1.17. At most, 20 prediction-filtering-excision iterations were performed.

The predicted retrotransposons were classified by their homology to the RepeatExplorer Viridiplantae 3.0 dataset<sup>38</sup> sequences. Their age was estimated under the assumption that their long terminal repeats (LTRs) were identical at the time of insertion by aligning their 3' and 5' LTRs using clustalo<sup>40,41</sup> 1.2.4, calculating their divergence (K) using the Kimura-2-parameter distance and dividing it by twice  $1.5 \times 10^{-8}$  substitution per site per year ( $\tau$ )<sup>42</sup>.

The predicted retrotransposons covered 26.6%, 32.2%, and 29.3% of the *N. sylvestris*, *N. tomentosiformis*, and *N. tabacum* genomes, respectively (Table 2). Regardless of the species, the most frequent element subclass is Ty3/gypsy|chromovirus|Tekay, representing between 40% and 56% of the total predicted retrotransposon length. The only element subclass that shows a marked difference between the three species is Ty3/gypsy|non-chromovirus|O TA|Tat|Ogre, which covers 116,167,517 bp (18.8% of the total predicted retrotransposon length) in *N. sylvestris*, and only 21,672,795 bp (3.9%) in *N. tomentosiformis*. In *N. tabacum*, it covers 135,653,424 bp (11.6%), close to the sum of its coverage in the two precursor species (137,840,312 bp). Looking at the predicted insertion ages, a recent expansion of the Alesia and Angela subclasses of Ty1/copia and of the Ogre subclass of Ty3/gypsy retrotransposons in *N. sylvestris* and *N. tabacum*, but not in *N. tomentosiformis*, is observed (Fig. 2).

**Coding-gene Prediction and Annotation.** Genomes were masked using blast<sup>43,44</sup> 2.14.0 windowmasker with dusting, and augustus<sup>45</sup> 3.5.0 was used for gene prediction. A training dataset was created by separately mapping *S. lycopersicum*, *S. tuberosum*, and *Nicotiana attenuata* cDNA and CDS from Ensembl 56 using minimap<sup>21,22</sup> 2.26 to the *N. sylvestris* and *N. tomentosiformis* genomes. Any sequence with an annotation matching 'hypothetical', 'unknown', 'polyprotein', 'domain-containing', 'chloroplast', or 'mitochondria' were omitted from the mapping. Gene models were constructed from the mapped sequences using bedtools<sup>46</sup> 2.30.0 and filtered using gffread<sup>47</sup> 0.12.7 with the parameters -V -H -U -N -P -J -M -K -Q -Y -Z -F --keep-exon-attrs. Training sequences were then extracted from the genomes using the obtained GFF annotation file and adding 1,000 bp

Metric	arabidopsis	tomato	coyote_tobacco	Nicotiana
Without hints				
Base_level_sensitivity	0.964	0.971	0.959	<b>0.976</b>
Base_level_specificity	0.887	0.917	<b>0.93</b>	0.929
Exon_level_sensitivity	0.841	0.857	0.822	<b>0.872</b>
Exon_level_specificity	0.731	0.802	0.812	<b>0.832</b>
Gene_level_sensitivity	0.335	0.408	0.371	<b>0.443</b>
Gene_level_specificity	0.29	0.369	0.367	<b>0.418</b>
UTR_nucleotide_level_sensitivity	<b>0.623</b>	0.475	0.434	0.475
UTR_nucleotide_level_specificity	0.455	0.487	0.492	<b>0.557</b>
UTR_exon_level_sensitivity	<b>0.183</b>	0.16	0.151	0.17
UTR_exon_level_specificity	0.162	0.159	0.177	<b>0.185</b>
Accuracy	0.745533	0.7844	0.771333	<b>0.804933</b>
With hints				
Base_level_sensitivity	0.987	0.991	0.991	<b>0.992</b>
Base_level_specificity	0.945	0.953	<b>0.965</b>	0.959
Exon_level_sensitivity	0.955	0.954	0.953	<b>0.956</b>
Exon_level_specificity	0.904	0.915	<b>0.926</b>	0.924
Gene_level_sensitivity	<b>0.751</b>	0.742	0.737	0.749
Gene_level_specificity	0.695	0.696	<b>0.719</b>	0.718
UTR_nucleotide_level_sensitivity	<b>0.598</b>	0.457	0.417	0.437
UTR_nucleotide_level_specificity	0.612	0.611	<b>0.706</b>	0.696
UTR_exon_level_sensitivity	<b>0.236</b>	0.21	0.216	0.227
UTR_exon_level_specificity	0.223	0.204	0.236	<b>0.238</b>
Accuracy	0.905333	0.908	0.9124	<b>0.913733</b>

**Table 3.** Augustus testing metrics with the arabidopsis, tomato, coyote\_tobacco, and Nicotiana models. The best scores are highlighted in bold. Accuracy is calculated as  $(3 \times \text{nsen} + 2 \times \text{nspe} + 4 \times \text{esen} + 3 \times \text{espe} + 2 \times \text{gsen} + 1 \times \text{gspe})/15$ .

flanking regions. One-fourth of the gene models were set aside for testing for each combination of species and dataset. After merging the training and testing datasets, a Nicotiana model was trained using the etraining and optimize\_augustus.pl programs bundled with augustus<sup>45</sup> 3.5.0. A total of 10,092 loci were used for training, and 3,362 loci were used for testing.

To hint at the augustus predictions, Ensembl 56 proteins from *S. lycopersicum*, *S. tuberosum*, and *N. attenuata* were mapped to the genomes using minipro<sup>48</sup> 0.11, and aletsch<sup>49</sup> 1.0.3 was used to construct transcripts from Illumina paired-end RNA-Seq reads from SRR11912457<sup>50</sup>, SRR2106531<sup>51</sup>, ERR274387<sup>52</sup>, ERR274388<sup>53</sup>, ERR274389<sup>54</sup>, ERR274390<sup>55</sup>, ERR274391<sup>56</sup>, ERR274392<sup>57</sup>, ERR274393<sup>58</sup>, ERR274394<sup>59</sup>, ERR274395<sup>60</sup>, ERR274396<sup>61</sup>, ERR274397<sup>62</sup>, ERR274398<sup>63</sup>, ERR274399<sup>64</sup>, ERR274400<sup>65</sup>, ERR274401<sup>66</sup>, ERR274402<sup>67</sup>, ERR274403<sup>68</sup>, ERR274404<sup>69</sup>, and ERR274405<sup>70</sup> mapped using hisat2<sup>34</sup> 2.2.1, and Oxford Nanopore long cDNA reads from SRR12045991<sup>71</sup>, SRR12045992<sup>72</sup>, SRR12045993<sup>73</sup>, and SRR12045994<sup>74</sup> mapped with minimap2<sup>21,22</sup> 2.26.

Augustus<sup>45</sup> 3.5.0 predictions were obtained using the trained Nicotiana model, the extrinsic.MPE.cfg extrinsic configuration file, and hints derived from the minipro<sup>48</sup> 0.11 and aletsch<sup>49</sup> 1.0.3 output with priorities of 4 and 3, respectively. Other augustus<sup>45</sup> 3.5.0 parameters used were --alternatives-from-evidence=off --alternatives-from-sampling=off --softmasking=1 --strand=both --genemodel=complete --UTR=on. Predicted gene models without supporting hints that did not encode a protein found in a uniprot eudicotyledons proteins dataset filtered to omit proteins with annotations matching 'uncharacterized', 'unknown', 'hypothetical', 'genome', 'domain-containing', 'family', 'transmembrane', 'putative', 'probable', 'predicted', 'member', 'fragment', 'truncated', 'superfamily', 'chloroplast', 'mitochond', 'low quality', or 'At.g' when using diamond<sup>39</sup> 2.1.6 blastx with the parameters --max-target-seqs 1 --min-score 200 --ultra-sensitive --frameshift 15 were removed.

To complement the augustus predictions, additional gene models were created by separately mapping the predicted *N. sylvestris*, *N. tomentosiformis*, and *N. tabacum* cDNA and CDS and the *S. lycopersicum*, *S. tuberosum*, and *N. attenuata* cDNA and CDS from Ensembl 56 to the genomes using minimap2<sup>21,22</sup> 2.26. Models that overlapped augustus predictions by 25% or more according to bedtools<sup>46</sup> 2.30.0 intersect were then filtered out by IDs using gffread<sup>47</sup> 0.12.7 with the parameters -P -M -K -Q -Y -Z -F, and the remaining genes models were added to those predicted with augustus<sup>45</sup> 3.5.0.

Functional annotation of the gene models was performed using diamond<sup>39</sup> 2.1.6 blastx with the parameters --max-target-seqs 1 --min-score 200 --ultra-sensitive --frameshift 15 and uniprot eudicotyledons proteins filtered to omit proteins with annotations matching 'uncharacterized', 'unknown', 'hypothetical', 'genome', 'domain-containing', 'family', 'transmembrane', 'putative', 'probable', 'predicted', 'member', 'fragment', 'truncated', 'superfamily', 'chloroplast', 'mitochond', 'low quality' or 'At.g'. Gene models overlapping with retrotransposons by 75% or more according to bedtools<sup>46</sup> 2.30.0 intersect and those with annotations matching 'transposon', 'transposase', 'polyprotein', 'gagpol', or 'gag-pol' were excluded to yield the final set of annotated gene models.

	counts			percent		
	Genome	Transcripts	Proteins	Genome	Transcripts	Proteins
<i>N. sylvestris</i>						
Complete BUSCOs (C)	5847	5657	5519	98.3%	95.1%	92.8%
Complete and single-copy BUSCOs (S)	5605	5434	5298	94.2%	91.3%	89.0%
Complete and duplicated BUSCOs (D)	242	223	221	4.1%	3.7%	3.7%
Fragmented BUSCOs (F)	8	97	144	0.1%	1.6%	2.4%
Missing BUSCOs (M)	95	196	287	1.6%	3.3%	4.8%
Total BUSCO groups searched	5950	5950	5950	100.0%	100.0%	100.0%
<i>N. tomentosiformis</i>						
Complete BUSCOs (C)	5858	5716	5560	98.5%	96.1%	93.4%
Complete and single-copy BUSCOs (S)	5660	5517	5351	95.1%	92.7%	89.9%
Complete and duplicated BUSCOs (D)	198	199	209	3.3%	3.3%	3.5%
Fragmented BUSCOs (F)	12	79	140	0.2%	1.3%	2.4%
Missing BUSCOs (M)	80	155	250	1.3%	2.6%	4.2%
Total BUSCO groups searched	5950	5950	5950	100.0%	100.0%	100.0%
<i>N. tabacum</i>						
Complete BUSCOs (C)	5901	5837	5774	99.2%	98.1%	97.0%
Complete and single-copy BUSCOs (S)	525	835	996	8.8%	14.0%	16.7%
Complete and duplicated BUSCOs (D)	5376	5002	4778	90.4%	84.1%	80.3%
Fragmented BUSCOs (F)	1	38	66	0.0%	0.6%	1.1%
Missing BUSCOs (M)	48	75	110	0.8%	1.3%	1.8%
Total BUSCO groups searched	5950	5950	5950	100.0%	100.0%	100.0%

**Table 4.** Statistics of the BUSCO genome, transcripts, and proteins completeness evaluation using the solanales\_odb10 lineage dataset for *Nicotiana sylvestris*, *Nicotiana tomentosiformis* and *Nicotiana tabacum*.

## Data Records

The genomes and annotations are available from Zenodo under records 8256252<sup>75</sup>, 8256254<sup>76</sup>, and 8256256<sup>77</sup>. The trained *Nicotiana* model for augustus gene prediction is available from Zenodo under record 8256280<sup>78</sup>.

The genomes have been deposited at DDBJ/ENA/GenBank under the accessions ASAF00000000<sup>79</sup>, ASAG00000000<sup>80</sup> and AWOJ00000000<sup>81</sup>.

Raw sequencing data are available from the National Center for Biotechnology Information Short Read Archive under accessions SRR25685126<sup>82</sup>, SRR25685127<sup>83</sup>, SRR25685128<sup>84</sup>, SRR25685129<sup>85</sup>, and SRR25685130<sup>86</sup> in BioProject PRJNA182500, SRR25685034<sup>87</sup>, SRR25685035<sup>88</sup>, SRR25685036<sup>89</sup>, SRR25685037<sup>90</sup>, SRR25685038<sup>91</sup>, SRR25685039<sup>92</sup>, and SRR25685040<sup>93</sup> in BioProject PRJNA182501, and SRR25685386<sup>94</sup>, SRR25685387<sup>95</sup>, SRR25685388<sup>96</sup>, SRR25685389<sup>97</sup>, SRR25685390<sup>98</sup>, SRR25685391<sup>99</sup>, SRR25685392<sup>100</sup>, SRR25685393<sup>101</sup>, SRR25685394<sup>102</sup>, SRR25685395<sup>103</sup>, and SRR25685396<sup>104</sup> in BioProject PRJNA208210 for *N. sylvestris*, *N. tomentosiformis*, and *N. tabacum*, respectively.

## Technical Validation

The quality and completeness of the assemblies were assessed with yak<sup>105</sup> 0.1 using 20% of the processed Illumina short-reads which were set aside for that purpose. For *N. tabacum*, Quality Coverage and Quality Value of 0.982 and 38.1 were obtained; for *N. sylvestris*, they were of 0.993 and 41.5; and for *N. tomentosiformis* they were of 0.991 and 43.2.

The quality of the gene predictions from the trained *Nicotiana* model was evaluated using the prepared testing sets and compared with results obtained using already available models for arabidopsis, tomato, and coyote\_tobacco models (Table 3).

The completeness of the gene model sets was evaluated using BUSCO<sup>106</sup> 5.4.7 with the solanales\_odb10 lineage dataset. Completeness of 98.1%, 95.1%, and 96.1% at the transcript level and of 97.0%, 92.8%, and 93.4% at the protein level were obtained for *N. tabacum*, *N. sylvestris*, and *N. tomentosiformis*, respectively (Table 4). These values are similar to those obtained for *S. lycopersicum*, of 95.0% at the transcript level and 92.3% at the protein level.

## Code availability

All software used in this work is publicly available, with versions and parameters clearly described in Methods. If no detailed parameters were mentioned for a software, the default parameters suggested by the developer were used. No custom code was used during this study for the curation and/or validation of the datasets.

Received: 9 November 2023; Accepted: 12 January 2024;

Published online: 26 January 2024

## References

- Knapp, S., Bohs, L., Nee, M. & Spooner, D. M. Solanaceae—A model for linking genomics with biodiversity. *Comp. Funct. Genomics* **5**, 285–291 (2004).
- Olmstead, R. G. *et al.* A molecular phylogeny of the Solanaceae. *Taxon* **57**, 1159–1181 (2008).
- Clarkson, J. J. *et al.* Phylogenetic relationships in *Nicotiana* (Solanaceae) inferred from multiple plastid DNA regions. *Mol. Phylogenet. Evol.* **33**, 75–90 (2004).
- Clarkson, J. J. *et al.* Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytol.* **168**, 241–252 (2005).
- D'Andrea, L. *et al.* Polyploid *Nicotiana* section *Suaveolentes* originated by hybridization of two ancestral *Nicotiana* clades. *Front. Plant Sci.* **14** (2023).
- Baldwin, I. T. Inducible Nicotine Production in Native *Nicotiana* as an Example of Adaptive Phenotypic Plasticity. *J. Chem. Ecol.* **25**, 3–30 (1999).
- Kaminski, K. P. *et al.* Alkaloid chemophenetics and transcriptomics of the *Nicotiana* genus. *Phytochemistry* **177**, 112424 (2020).
- Tissier, A. Trichome Specific Expression: Promoters and Their Applications. in *Transgenic Plants - Advances and Limitations* (InTech, 2012).
- Sierro, N. *et al.* Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol.* **14**, R60 (2013).
- Sierro, N. *et al.* The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* **5**, (2014).
- Edwards, K. D. *et al.* A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics* **18**, (2017).
- NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274527> (2013).
- NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274528> (2013).
- NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274540> (2013).
- NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274542> (2013).
- Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
- Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
- Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, (2023).
- Mak, Q. X. C., Wick, R. R., Holt, J. M. & Wang, J. R. Polishing De Novo nanopore assemblies of bacteria and eukaryotes with FMLRC2. *Mol. Biol. Evol.* **40**, (2023).
- Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, (2018).
- Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
- Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
- Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
- Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
- Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing <https://doi.org/10.48550/ARXIV.1207.3907> (2012).
- Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).
- NCBI Genome Project. *Nicotiana tabacum* plastid, complete genome. *Nucleotide* [https://identifiers.org/nucleotide/NC\\_001879.2](https://identifiers.org/nucleotide/NC_001879.2) (2000).
- NCBI Genome Project. *Nicotiana tabacum* mitochondrion, complete genome. *Nucleotide* [https://identifiers.org/nucleotide/NC\\_006581.1](https://identifiers.org/nucleotide/NC_006581.1) (2004).
- Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
- Langmead, B. Kraken 2, KrakenUniq and Bracken indexes <https://benlangmead.github.io/aws-indexes/k2> (2022).
- Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
- High Performance Algorithms Group. The Wellcome Sanger Institute. Paired REad TEXTure Mapper <https://github.com/wtsi-hpag/PretextMap> (2022).
- High Performance Algorithms Group. The Wellcome Sanger Institute. OpenGL Powered Pretext Contact Map Viewer <https://github.com/wtsi-hpag/PretextView> (2022).
- Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
- Bindler, G. *et al.* A high density genetic map of tobacco (*Nicotiana tabacum* L.) obtained from large scale microsatellite marker development. *Züchter Genet. Breed. Res.* **123**, 219–230 (2011).
- Wu, F. & Tanksley, S. D. Chromosomal evolution in the plant family Solanaceae. *BMC Genomics* **11**, 182 (2010).
- Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, (2008).
- Neumann, P., Novák, P., Hošťáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **10**, (2019).
- Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
- Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences: Clustal Omega for Many Protein Sequences. *Protein Sci.* **27**, 135–145 (2018).
- Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, (2011).
- Mokhtar, M. M., Alsamman, A. M. & El Allali, A. PlantLTRdb: An interactive database for 195 plant species LTR-retrotransposons. *Front. Plant Sci.* **14**, (2023).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
- Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- Perteza, G. & Perteza, M. GFF utilities: GffRead and GffCompare. *F1000Res.* **9**, 304 (2020).
- Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
- Shao, M. Assembler for multiple RNA-seq samples <https://github.com/Shao-Group/aletsch> (2020).
- NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR11912457> (2020).
- NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR2106531> (2016).



52. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274387> (2013).
53. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274388> (2013).
54. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274389> (2013).
55. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274390> (2013).
56. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274391> (2013).
57. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274392> (2013).
58. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274393> (2013).
59. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274394> (2013).
60. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274395> (2013).
61. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274396> (2013).
62. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274397> (2013).
63. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274398> (2013).
64. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274399> (2013).
65. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274400> (2013).
66. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274401> (2013).
67. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274402> (2013).
68. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274403> (2013).
69. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274404> (2013).
70. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:ERR274405> (2013).
71. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR12045991> (2021).
72. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR12045992> (2021).
73. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR12045993> (2021).
74. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR12045994> (2021).
75. Sierrro, N. *Nicotiana sylvestris* genome assembly and annotation. *Zenodo* <https://doi.org/10.5281/zenodo.8256252> (2023).
76. Sierrro, N. *Nicotiana tomentosiformis* genome assembly and annotation. *Zenodo* <https://doi.org/10.5281/zenodo.8256254> (2023).
77. Sierrro, N. *Nicotiana tabacum* genome assembly and annotation. *Zenodo* <https://doi.org/10.5281/zenodo.8256256> (2023).
78. Sierrro, N. *Nicotiana* model for augustus gene prediction, *Zenodo*, <https://doi.org/10.5281/zenodo.8256280> (2023).
79. Sierrro, N. & Ivanov, N. V. *Nicotiana sylvestris*, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:ASAF00000000> (2023).
80. Sierrro, N. & Ivanov, N. V. *Nicotiana tomentosiformis*, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:ASAG00000000> (2023).
81. Sierrro, N. & Ivanov, N. V. *Nicotiana tabacum* cultivar K326, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:AWOJ00000000> (2023).
82. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685126> (2023).
83. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685127> (2023).
84. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685128> (2023).
85. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685129> (2023).
86. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685130> (2023).
87. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685034> (2023).
88. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685035> (2023).
89. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685036> (2023).
90. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685037> (2023).
91. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685038> (2023).
92. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685039> (2023).
93. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685040> (2023).
94. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685386> (2023).
95. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685387> (2023).
96. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685388> (2023).
97. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685389> (2023).
98. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685390> (2023).
99. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685391> (2023).
100. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685392> (2023).
101. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685393> (2023).
102. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685394> (2023).
103. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685395> (2023).
104. NCBI Sequence Read Archive, <https://identifiers.org/ncbi/insdc.sra:SRR25685396> (2023).
105. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
106. Manni, M., Berkeley, M. R., Seppely, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

## Acknowledgements

We thank Simon Goepfert and Nicolas Bakaher for scientific discussions, and Rebecca Higgins for manuscript editorial revision.

## Author contributions

N.S. and N.V.I. conceived this project; M.A. and R.D. performed the experiments; N.S. assembled the genomes, generated the annotation sets, and performed the data analysis; N.S. and N.V.I. wrote and revised the manuscript. All authors have read and approved the final manuscript.

## Competing interests

N.S., M.A., R.D., and N.V.I. are employees of Philip Morris International.

## Additional information

**Correspondence** and requests for materials should be addressed to N.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024