Document downloaded from:

https://riunet.upv.es/handle/10251/236182

This paper must be cited as:

Ranawaka, B.; An, J.; Lorenc, MT.; Jung, H.; Sulli, M.; Aprea, G.; Roden, S.... (2023). A multi-omic Nicotiana benthamiana resource for fundamental research and biotechnology. Nature Plants (Online). 1558-1571. https://doi.org/10.1038/s41477-023-01489-8

The final publication is available at

https://doi.org/10.1038/s41477-023-01489-8

Copyright Nature Publishing Group

Additional Information

Editor summary:

Chromosome-level genome reference sequence assemblies of the model and biofactory Nicotiana benthamiana line, and a wild relative, have been generated and annotated for gene models, tissue-specific transcriptomes, microRNAs and epigenetic landscapes.

Reviewer Recognition:

Nature Plants thanks Yongbiao Xue, Ed Rybicki and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Figure or Table # Please group Extended Data items by type, in sequential order. Total number of items (Figs. + Tables) must not exceed 10.	Figure/Table title One sentence only	Filename Whole original file name including extension. i.e.: Smith_ED_Fig1.jpg	Figure/Table Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	Profiles of average emission of selected putative insect-attracting volatile compounds and nicotine	Figure E1_300dpi.jpg	Profiles of average emission of selected putative insect-attracting volatile compounds and nicotine (a defence compound) in green leaf and floral headspace of LAB and QLD over a 24-hr period. (**A**) LAB floral headspace (**B**) QLD floral headspace (**C**) LAB green leaf headspace (**D**) QLD green leaf headspace. These results indicate that QLD flowers, but not LAB flowers or LAB and QLD leaves, emit benzyl alcohol overnight (6:00 pm - 8:00 am). Error bars represent the standard error of the mean (n=4 per sample point).
Extended Data Fig. 2	Differentially accumulated metabolites in semi-polar extracts of tissues from *N. benthamiana* LAB and QLD	Figure E2_300dpi.jpg	Differentially accumulated metabolites in semi-polar extracts of *N. benthamiana* LAB vs QLD tissues analysed by liquid chromatography/high resolution mass spectrometry (LC/HESI/MS). The degree of orange/blue indicates relative levels in LAB vs QLD, grey shaded areas not detectable levels
Extended Data Fig. 3	Cladogram of relationships of the nicotine demethylase genes in *S. lycopersicum N. sylvestris, N. tabacum, N. tomentosiformis, N. attenuate,* and *N. benthamiana* (LAB and QLD)	Figure E3_300dpi.jpg	Cladogram of relationships of the nicotine demethylase genes in *S. lycopersicum, N. sylvestris, N. tabacum, N. tomentosiformis, N. attenuata,* and *N. benthamiana* (LAB and QLD). The highlighted clade contains the *N. benthamiana* CYP82E2 gene. Genes without stars represent proteins of uncharacterized nicotine N-demethylase activity. (B) Location of bZIP transcription factor binding motifs (red and purple triangles) in LAB and QLD 2kb promoter. The bottom panel shows the transversion in the third TF binding motif (purple triangles) that probably inhibits TF binding and expression of CYP82E2 in LAB. (C) Gene expression (TPM) of CYP82E2

			in leaf and flower tissues of LAB and QLD. Error bars represent the standard error of the mean (n=3 biologically independent flower and leaf samples of LAB and QLD)
Extended Data Fig. 4	(A).Plot of contact matrices of LAB and QLD assemblies (B).Synteny of Self-incompatibility (S)-like loci in tomato, *N. attenuata*, *N.tabacum*, petunia, LAB and QLD, cladogram of gene sequence similarities and tissue-expression of mRNA LAB S-locus genes	Figure E4_300dpi.jpg	**(A**) Plot of contact matrices of LAB and QLD assemblies. Juicebox plot from HiC analysis showing resolution into 19 contiguous elements (chromosomes) for both LAB and QLD assemblies. **(B)** Synteny of self-incompatibility (S)-like loci in tomato, *N. attenuata*, *N.tabacum*, petunia, LAB and QLD, cladogram of protein gene sequence similarities and tissue-specific mRNA expression of the LAB S-locus. Gene arrangement and relationships in cartoon form of the genes in the highly recombinogenic S-locus (comprised of an S-RNAse and associated multiple copies of F-box (SLF) proteins) in the most advanced genome assemblies of tomato, *N. attenuata*, *N.tabacum*, petunia, LAB and QLD. The colours of the genes represent their relationships across species, as indicated in the cladogram. The analysis shows contiguity of the S-locus in tomato, LAB and QLD and the fragmented nature of the locus in *N. attenuata*, *Petunia axillaris*, due to their presence on small scaffolds, and the incomplete assembly of Ch22 in *N.tabacum*. Tissue expression data for LAB shows that the intervening gene 16g24630 is expressed in all 5 tissues examined but the S-RNAse and SLF genes are expressed only in the floral tissue, as expected for a floral incompatability-associated locus. Distances between genes are indicated in Mb
Extended Data Fig. 5	miRNA families in LAB and QLD shared with *A. attenuata*, *S. lycopersicum*, and *S. tuberosum*	Figure E5_300dpi.jpg	The number of identified miRNA families in LAB and QLD that are shared with three Solanaceae plants (*A. attenuata*, *S. lycopersicum*, and *S. tuberosum*) and the well-studied plant Arabidopsis (*A. thaliana*) are illustrated in a Venn diagram. The figure shows that the major miRNAs in the most related plant, *N. attenuata*, were identified in both LAB and QLD. Many potential miRNAs were discovered that have not been

			previously identified. Subfigure (A) shows the overlapping number of identified miRNAs in LAB that are shared with the other four species. Subfigure (B) shows the identified miRNAs in QLD
Extended Data Fig. 6	Transformation efficiencies of LAB, QLD and Northern Territory (NT) accessions	Figure E6_300dpi.jpg	Comparison of transformation efficiencies of LAB, QLD and Northern Territory (NT) accessions. (**A)** Regeneration, selection, shoot development, and root development of LAB, NT and QLD ecotypes post-transformation with a 35S:Cas9 cassette and kanamycin selectable marker (scale bar represents 1 cm). The dates on top of the image indicate the progression of transformation. (**B**) Comparison of time taken for regeneration, growth (1-2 cm shoots) and rooting of LAB, QLD and NT. ANOVA two-tailed test was performed to determine the significance differences. (Data are presented as mean values +/- standard error (n=3 biologically independent samples). (**C**) Comparison of regeneration frequency and transformation efficiency of LAB, QLD and NT. ANOVA two-tailed test without transformation was performed to determine the significance differences between percentage data derived from count data. Independent positive transformants of LAB n=72, QLD n=74 and NT n=21 (a single sister plant derived from one single callus) were used to calculate the transformation efficiency.
Extended Data Fig. 7	Comparison of CRISPR/Cas9 editing efficiency in LAB and QLD	Figure E7_300dpi.jpg	Comparison of CRISPR/Cas9 editing efficiency in LAB and QLD. (**A**) The basic editing construct (with kanamycin selection) used to transform LAB or QLD tissues. The two guide (g)RNA sequences were placed between the tRNA processing units (indicated as spacer sequences 1 & 2 in panel A). Two sites were chosen within the same target gene, usually ~200 nucleotides apart, and gave either a dropout of the intervening DNA sequence in the genome or inaccurate repair of one or both sites. (**B**) Phenotypes of QLD knockouts (ko) of RDR1 infected with Tobacco mosaic virus (TMV), RDR6 and Phytoene desaturase (PDS) and

			LAB knockout of RDR2. Silencing of PDS in QLD targeted two homoeologs simultaneously to give biallelic silencing of both genes in the T0 generation. gRNA sequences used: *RNA-dependent RNA polymerase* (NbRDR1): TAAATAGTACAGTTTCTCCA; GACACTCAAAGTTTCTCTGG NbRDR2: CCACTCCCAACGTAGATAAG; GTGTCTCGAAATGTGCTGCA NbRDR6: CTTACTTAGAAGTCATCAGG; CTGCAACAGTATTACCAAAG *Phytoene desaturase* (NbPDS) TCACAAACCGATATTGCTGG; GAGCTTCAGGAAAATCAAAG (**C**) Comparison of editing efficiency of LAB and QLD. Editing efficiency in LAB and QLD was determined using the NbRDR genes involved in RNAi.
Extended Data Fig. 8	Comparison of ERF locus IX and AN-like MYB loci in LAB and QLD with other Solanaceae	Figure E8_300dpi.jpg	**(A1)** Synteny analysis of the ERF locus IX in tomato, *N. obtusifolia*, LAB and QLD shows lineage-specific tandem duplications of ERF189s, advanced diploidization through loss of gene function, and an inversion between LAB and QLD on chr 14, flanked by newly inserted Copia elements. Functional genes are shown in green; nonfunctional/pseudogenes are in blue. Gypsy, Copia and LTRs are indicated as yellow, olive green and red arrows respectively. Shading indicates the orthology relationships of ERF189 genes between different syntenic blocks. The inverted region of LAB chromosome 14 and the Gypsy and Copia landscape within the blue box is magnified in the second panel (**A2**). The third panel (**A3**) is further magnifying the region indicated by a red box in (A2). The fourth panel (**A4**) depicts the epigenetic landscape (H3K4me3, H3K9me2 and cytosine methylation) and the expression of selected ERF189 genes in LAB. For H3K4me3, H3K9me2 enriched regions are shown in blue and the lack of histone modification is in red. Methylated cytosines are shown

			as blue bars. (**A5**) Tissue-specific gene expression of Ancestral (the left-most and right-most two genes indicated in green in *N. obtusifolia*) and "Expansion" (the three green genes in the middle of *N. obtusifolia*) genes. **(B1)** Synteny analysis of the AN-like locus in tomato, LAB and QLD shows tandem duplication of SlAN2-like MYB genes in LAB and QLD with loss of gene function of 1 copy in QLD (Bur1) and both copies (Bur 1 & 2) in LAB. Loss of Bur2 in LAB is associated with a newly inserted Copia element. Functional genes are shown in bright green, and nonfunctional/pseudogenes are in dark green. Gypsy, Copia and LTRs are indicated as yellow, olive green and red arrows respectively. Shading indicates the orthology relationships. The Gypsy and Copia landscape within the blue box are zoomed in the second panel (**B2**) The third panel (**B3**) shows the amino acid change in LAB Bur2 which alters its bHLH binding site. (**B4**) shows the function of Bur1 is defective in LAB, QLD and NT, and that Bur2 is fully active in QLD and NT and may be partially restored by simultaneous overexpression of bHLH in LAB. Bur3 is only functional in NT. (**B5**) Levels of different anthocyanins in LAB and QLD leaves following transient expression of AcMYB110 (an AN-like MYB from Kiwifruit) or QLD Bur2. For comparison, the Anthocyanin levels were measured in NT stably transformed with an AcMYB110 construct.
Extended Data Fig. 9	Comparison of RPM1-like locus and Cytochrome P450 loci in LAB and QLD with other Solanaceae	Figure E9_300dpi.jpg	**(A)** Synteny of RPM1-like loci in tomato, *N. attenuata*, *N.tabacum*, LAB and QLD. **(B)** Synteny of a terpene biosynthesis pathway Cytochrome P450 locus in *N. attenuata*, LAB and QLD. Gene arrangement in cartoon form representing RPM1-like bacterial resistance genes and CYP736A-like genes (functional - bright green), possibly functional (dark green), defective/pseudogenes (blue). In (A), distances between genes indicated (black text)< 15kbp; (red

			text) >15kbp and surrounding syntenic genes in are shown in orange, purple, yellow and brown. Orthology/homology relationships are indicated by coloured shading. In (B), distances between genes indicated (black text)< 50kbp; (red text) >50kbp. TE annotation tracks for LAB and QLD were prepared using annotation data from the EDTA TE annotation pipeline (see online methods) and Geneious Prime software (Geneious Prime® 2023.0.1; https://www.geneious.com). Only LTR-transposable elements are shown. Yellow blocks represent GYPSY elements and green blocks represent COPIA elements. The size of each block is proportional to the number of base-pairs annotated for that element. Red triangles represent LTR repeat regions that flank either a GYPSY or COPIA element. These elements are likely to be nearly complete and can be considered possible autonomous elements. The rectangular red blocks flank unknown LTR-TE elements. Unknown TEs are elements that are recognized as an LTR element but are not able to be classified as either a COPIA or GYPSY element due to irregularities in internal sequences for that element. These are likely to represent non-autonomous elements. Those elements not flanked by LTR sequences are highly fragmented non-functional elements. The blue rectangular boxes highlight the location of the genes annotated in the tracks above and below the TE annotation tracks

13

Item	Present?	Filename Whole original file name including extension. i.e.: Smith_SI.pdf. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_Figures.pdf	Figure S1. Schematic of anthocyanin biosynthesis pathway showing genes upregulated in N.

			benthamiana after agroinfiltration with an AN-like MYB construct. Figure S2. Schematic of assembly and annotation pipelines of the N. benthamiana genomes. Figure S3. Upset plot showing orthologous groups among LAB, QLD, N. tabacum, N.sylvestris, N. tomentosiformis, N. glauca, A. thaliana, V. vinifera, S. lycopersicum and S. tuberosum. Figure S4. Histogram showing completeness and quality of the LAB and QLD annotations based on predicted protein lengths relative to orthologs in Arabidopsis. Figure S5. SynVisio waterfall plots showing the syntenic relationships between chromosomes of the LAB subgenomes and those of the N. sylvestris derived subgenome of tobacco. Figure S6. Average relative homeolog expression in subgenomes of N. benthamiana. Figure S7. RNAi associated genes Figure S8. RDR knockout figures Figure S9. Copia element density and methylation profiles of chromosomal regions in the proximity of genes in LAB. Figure S10. Association of genomic features with T-DNA genomic junctions. Figure S11. Distances to the closest gene for insertion sites for transgenes and intact Copia and Gypsy insertion sites. Figure S12. Box and whisker plot of average intergenic distances of LAB, QLD, tomato and N.attenuata genomes. Figure S13. Inter-fertility of LAB and QLD.
Reporting Summary	Yes	Reporting_summary.pdf	

14

Type	Number	Filename	Legend or Descriptive Caption Describe the contents of the file

	Each type of file (Table, Video, etc.) should be numbered from 1 onwards. Multiple files of the same type should be listed in sequence, i.e.: Supplementary Video 1, Supplementary Video 2, etc.	Whole original file name including extension. i.e.: *Smith_ Supplementary_Video_1.mov*	
Supplementary Table	1	Supplementary_Tables.xlsx	Table S1 Morphological, developmental, and metabolic differences between LAB and QLD. Table S2 Identification of differentially expressed semi-polar metabolites in LAB vs QLD. Table S3 Size and number of genes in each chromosome in the LAB and QLD assemblies Table S4 All expressed genes in LAB Table S5 All expressed genes in QLD Table S6 Quality statistics of the LAB and QLD assemblies Table S7A Mapping genes in specific solanacae species with LAB to find % of genes in LAB with orthologs in these species Table S7B Protein cluster analysis Table S7C Characteristics and sources of genomes used for comparative analyses Table S8A Expressed miRNAs in LAB Table S8B Expressed miRNAs in QLD Table S8C Full list of potential miRNAs in LAB Table S8D Full list of potential miRNAs in LAB Table S8E Plant species used in miRPlant analyses Table S9 SNPs between different N. benthamiana sequenced accessions. Table S10 Number of homeologous genes among chromosomes of LAB, N. attenuata, and the N. sylvestris and N. tomentosiformis subgenomes of N. tabacum

			Table S11A Distribution of LAB homeologous genes across chromosomes of the subgenomes Table S11B Distribution of QLD homeologous genes across chromosomes of the subgenomes Table S11C Homoeologs present on partner subgenomes in LAB Table S11D Homoeologs present on partner subgenome but not partner chromosomes in LAB Table S11E Homoeologs present on same subgenome in LAB Table S11F Homoeologs present on partner subgenome in QLD Table S11G Homoeologs present on partner subgenome but not partner chromosomes in QLD Table S11H Homoeologs present on same subgenome in QLD Table S12 Homeolog expression pattern detection for subgenome dominance evaluation

15

Parent Figure or Table	Filename Whole original file name including extension. i.e.: *Smith_SourceData_Fig1.xls*, or *Smith_ Unmodified_Gels_Fig1.pdf*	Data description i.e.: Unprocessed western Blots and/or gels, Statistical Source Data, etc.
Source Data Fig. 4	**gels original.pdf**	Unprocessed SDS-PAGE

16

A multi-omic *Nicotiana benthamiana* resource for fundamental research and biotechnology

Buddhini Ranawaka[1,2*], Jiyuan An[1,2*], Michał T. Lorenc[1], Hyungtaek Jung[1,3], Maria Sulli[4], Giuseppe Aprea[4], Sally Roden[1,2], Victor Llaca[5], Satomi Hayashi[1,2], Leila Asadyar[1,2], Zacharie LeBlanc[1], Zuba Ahmed[1,2], Fatima Naim[1,6], Samanta Bolzan De Campos[1], Tal Cooper[1], Felipe F. de Felippes[1], Pengfei Dong[7], Silin Zhong[7], Victor Garcia-Carpintero[8], Diego Orzaez[8], Kevin J. Dudley[9,10], Aureliano Bombarely[8,11], Julia Bally[1,2], Christopher Winefield[2,12], Giovanni Giuliano[4] & Peter M. Waterhouse[1,2]

[1] Centre for Agriculture and the Bioeconomy, Queensland University of Technology (QUT), Brisbane, QLD 4001, Australia
[2] ARC Centre of Excellence for Plant Success in Nature & Agriculture, Brisbane, QLD 4001, Australia
[3] Current address : Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation (QAAFI), The University of Queensland, Brisbane QLD 4072 Australia.
[4] Italian National Agency for New technologies, Energy and Sustainable Economic Development (ENEA), Casaccia Res Ctr, Via Anguillarese 301, 00123 Roma Italy.
[5] Genomics Technologies, Corteva Agriscience, Johnston, IA 50131, USA.
[6] Current address: Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Bentley, WA 6102, Australia.
[7] State Key Laboratory of Agrobiotechnology, School of Life Sciences, The Chinese University of Hong Kong, Hong Kong, China.
[8] Instituto de Biología Molecular y Celular de Plantas (IBMCP), Consejo Superior de Investigaciones Científicas (CSIC), Universidad Politècnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain.
[9] School of Biology and Environmental Science, Queensland University of Technology (QUT), Brisbane, QLD 4001, Australia.
[10] QUT Central Analytical Research Facility, Queensland University of Technology (QUT), Brisbane, QLD 4001, Australia.
[11] Università degli Studi di Milano, Via Celoria 26, 20133 Milano, Italy.
[12] Department of Wine Food and Molecular Biosciences, Faculty of Agriculture and Life Sciences, Lincoln University, PO Box 85054, Lincoln 7647, Canterbury, New Zealand

*Contributed equally to this manuscript
peter.waterhouse@qut.edu.au , chris.winefield@lincoln.ac.nz , j.an@qut.edu.au

Abstract

Nicotiana benthamiana is an invaluable model plant and biotechnology platform with a ~3Gb allotetraploid genome. To further improve its usefulness and versatility, we have produced high quality chromosome-level genome assemblies, coupled with transcriptome, epigenome, microRNA and transposable element datasets, for the ubiquitously used LAB strain and a related wild accession, QLD. Additionally, single nucleotide polymorphism (SNP) maps have been produced for a further two

laboratory strains and four wild accessions. Despite the loss of 5 chromosomes from the ancestral tetraploid, expansion of intergenic regions, widespread segmental allopolyploidy, advanced diploidisation, and evidence of recent bursts of Copia pseudovirus (Copia) mobility not seen in other *Nicotiana* genomes, the two subgenomes of *N.benthamiana* show large regions of synteny across the Solanaceae. LAB and QLD have many genetic, metabolic and phenotypic differences, including disparate RNAi responses, but are highly inter-fertile and amenable to genome editing and both transient and stable transformation. The LAB/QLD combination has the potential to be as useful as the Columbia-0/Landsberg errecta partnership, utilised from the early pioneering days of Arabidopsis genomics to today.

The genus *Nicotiana*, comprising ~75 species, is predominantly endemic to the Americas and Australia[1]. Like most *Solanaceae*, it has a basic chromosome number of 12, with haploid DNA content ranging from 1.37Gb to 6.27Gb[2]. Section *Suaveolentes* (nicely smelling) includes *N. benthamiana* and is the largest allotetraploid group in the genus (~35 species) with chromosome numbers ranging from 15 to 24, diagnostic of an allotetraplodization event followed by chromosome loss[3,4,5] (Figure 1A). Almost all species in this section are indigenous to Australasia, which they apparently colonized during the Pliocene transition ~5-6 MYA. The diploid ancestors of *N. benthamiana* most likely belonged to the *Sylvestres* and *Noctiflorae* sections, whose closest sequenced extant relatives are *N. sylvestris* (~2.6Gb) and *N. glauca* (~3.2Gb)[6–11], respectively.

N. benthamiana is a very important plant platform for biopharmaceutical protein and vaccine production[7,12] and has been instrumental for fundamental discoveries in RNA interference (RNAi), plant–pathogen interactions, metabolic pathway engineering, functional genomics, synthetic biology and gene editing[13]. All of this work has relied on plants derived from one accession that we term LAB, which appears to have originated from a single collection near the Granites gold mine in central Australia[7,14,15] (Figure 1B). Several additional accessions have recently been described [7,14–16].

In this paper, we report whole genome, epigenome and metabolome information for the LAB strain and the wild QLD accession, coupled with SNP maps for further laboratory and wild accessions. We examine their relationships across the Solanaceae and seek to understand both the evolutionary forces at play and the basis of LAB's amenability as a research tool.

Results

Additional *N. benthamiana* accession resource

The QLD wild accession exhibits many morphological, developmental and metabolic differences to LAB[7,14–16], such as outcrossing flowers[7,14–16], floral scent production at night, and the robust capacity to produce anthocyanins (Figures 1C-D, E1, S1, Table S1). Most notably, QLD is much less susceptible to viruses than LAB, which has been associated with a difference in RNA interference (RNAi) competence[7,14]. The levels of a range of metabolites such as phenolic acids, flavonoids, amino acid derivatives, and metabolites involved in defence responses[17–20], such as nornicotine and hydroxygeranyl-linalool diterpene glycosides (HGL-DTG), exhibit marked differences between LAB and QLD (Figures 1E-F, E2-3 Table S2). LAB exhibited a higher number of underexpressed/non-functional biosynthetic pathways than QLD, except for phenolic acids and HGL-DTGs. Because of these and potentially many more differential characteristics, their genetic distance (Figure 1A), and particularly their differences in viral defence capacity, both LAB and QLD were chosen for chromosome level genome sequence assemblies.

Genome assembly, annotation and genetic diversity.

Long and short sequence reads of the LAB and QLD accessions were assembled into 19 chromosomes for each genome (Online methods, Figure S2). The chromosomes ranged in size from 128 to 182 Mb, with total genome sizes of ~2.8Gb (LAB) and ~2.9Gb (QLD), of which 99% and 96% respectively anchored to chromosomes (Table S3). This represents ~94% of the expected genome size estimated from cytological staining[2]. The assemblies were annotated (Online methods, Figure S2) to 45,797 and 49,636 gene models in LAB and QLD (Table S3) respectively. Approximately 87% of the gene models in LAB and 75% in QLD are fully supported by RNA-sequencing (Table S4-5) and 98% of LAB EST sequences[21–23] mapped to the LAB genome CDS. According to several quality scores, including the LTR Assembly Index[24], the LAB and QLD assemblies were well above the standard requirements of the Earth Biogenome Project[25,26] (Table S6). They have higher contiguity than any published *Nicotiana* genome assemblies (Table 1); this is further illustrated by the contact matrices (E4A) and analysis of the well-studied *S* locus (E4B).

Gene mapping (Table S7A) revealed that 72%, 92% and 89% of the *N. benthamiana* genes are orthologous to those in tomato, *N. attenuata*, and tobacco, respectively. Similar numbers were obtained by protein cluster analysis (Figure S3, Table S7B). There were ~1,000 and ~3,000 genes specific to LAB and QLD, respectively. Based on BUSCO scores and comparison of the predicted protein lengths with their *Arabidopsis* best hits, the LAB and QLD annotations are better than most *Nicotiana* and *Solanaceae*

annotations (Table S7C, Figure S4). A total of 369 and 383 potential miRNA families, and the expression of 59 and 57 of them were detected in LAB and QLD, respectively (Tables S8A-E, Figure E5).

The previously described NT, SA, WA and NWA wild accessions[14] (Figure 1B), as well as the extensively used GFP-expressing transgenic line (16c) produced in David Baulcombe's laboratory[23,27] (EU-LAB) and (USA-LAB) were re-sequenced and mapped onto the LAB and QLD assemblies. Single Nucleotide Polymorphisms (SNP) frequencies[28] (Table S9) were very low amongst the three LAB accessions (<25 SNPs/Mb), showing that our LAB assembly is a tremendous resource for worldwide *N. benthamiana* laboratory isolates; SNPs between the four wild accessions mirrored the previously calculated evolutionary relationships[14] (Table S9) and were similar in range to those of 20 *Capsicum annuum* accessions[29]. SA and LAB, originally collected from geographically well separated locations, have close genetic similarity (~51 SNPs/Mb). One possible explanation is that Pitjuri (a chewing tobacco mixture often containing dried *N. benthamiana* aerial tissue) exchanged along ancient aboriginal traditional trading routes (Figure 1B) has transported seed between these locations over the last 60,000 years. The annotated genomes of LAB and QLD, containing tracks describing gene models, SNPs with other *N. benthamiana* isolates, gene expression across five tissues, location and expression of pre-miRNAs, and the epigenetic landscapes, are available on an interactive WebApollo browser[30] (https://apollo.nbenth.com/).

Homeologous chromosomes, subgenomes and chromosome loss

The genomes of most diploid Solanaceous species consist of 12 chromosome pairs (2x=2n=24) encoding about 35,000 genes[31]. *N. tabacum*, an allotetraploid formed about 0.2-0.4 MYA[8,9] has 24 chromosome pairs (2n=4x=48) encoding ~70,000 genes[32,33]. In the estimated 5-6MY since the hybridization event basal to the Australian *Nicotiana* clade, *N. benthamiana* has lost 5 chromosome pairs to give a genome of 2n=4x=38 (Figure 1A)[4,5] .

A mapping approach, similar to that used to identify the subgenomic memberships of the *N. tabacum* chromosomes[32–34], was applied to *N. benthamiana* and *N. tabacum* using sequences from the genomes of *N. sylvestris*, *N. glauca* and *N. tomentosiformis*. This recapitulated the previous tobacco results but, as previously predicted[8,9], did not differentiate the *N. benthamiana* chromosomes into a *N. glauca*- and a *N. sylvestris*-related subgenome (Figure 2A). Therefore, we took a different approach. Syntenic sequences

and blocks of orthologous genes were compared both within the highly syntenic LAB and QLD genomes, with *N. tabacum*[32] and *N. attenuata* genome assemblies[34] (Figure 2B). A dendrogram, derived from matrices of degrees of similarity of counterpart gene sequences of the *Nicotiana* set, clearly identified eight homeologous chromosome pairs and 3 orphan chromosomes (Figure 2C, Table S10).

To separate the genome into two functional subgenomes we took a disjoint subset partitioning approach, enabled by the ~50% of genes for which homeologous gene pairs were identified to be on chromosomes other than their predicted homeologous counterpart. Every combination of LAB chromosomes was assigned to two disjoint subsets and measured for the number of homeologous gene pairs distributed 1:1 between the two subsets. The best combination, excluding the genes on the three orphan chromosomes, gave a distribution of 8,543 gene pairs in opposite subgenomes and 1,999 gene pairs in the same subgenome (Tables S11A-H, Figure 2D). Visual comparison of *N. benthamiana* subgenomes with genomes of six other Solanaceous species using SynVisio[35] revealed remarkable long range synteny across the family, which was even more apparent as the percentage of genes on each chromosome of the species that are orthologous to those on each tomato chromosome, especially in chromosomes 1, 2, 3 and 4, but still discernible in *N. tabacum* up to chr 7 (Figure 3A-B). By contrast, in *N. benthamiana* this conservation declines rapidly after chromosome 4 (Figure 3B, 3E), probably due to the high degree of chromosomal rearrangements specific to this allopolyploid species.

The blocks of synteny between the two subgenomes of *N. benthamiana* are more numerous, larger and contiguous than with the *N. sylvestris*-derived subgenome of *N. tabacum* (Figure S5). To investigate this further, a cluster analysis was made using the proteomes predicted from our LAB assembly and the available scaffold assemblies of *N. sylvestris* and *N. glauca* (Figure 3C). The LAB genes identified as clustering with *N. sylvestris* but not *N. glauca* genes, and vice-versa, were mapped onto the LAB genome (Figure 3D). This revealed that, even in the gene-rich, large, Solanaceae-wide syntenic blocks, extensive recombination has occurred between the two ancestral subgenomes and suggests that the present *N. benthamiana* genome is the result of extensive "duplication/deletion" homeologous recombination[36], or of repeated hybridization amongst the derivative populations from the original allotetraploid *Nicotiana* at the base of the Suaveolentes. These processes have produced chromosomes composed of genes from both ancestral parents, explaining the greater synteny between *N. benthamiana's* homeologous chromosomes than with their *N. sylvestris* counterparts. This is also the likely cause of the low level of subgenome dominance (Figure S6, Table S12). Subgenomes A and B encode 23,408 and 22,388 genes,

178 respectively, and the overall transcript abundance of homeologs differs by only 1%, suggesting that the
179 genome is in balanced but fluid harmony.

180 **LAB and QLD as model plants and biofactory platforms**

181 An impaired RNAi response in *N.benthamiana*-LAB may underlie the plant's excellence as a biofactory and
182 research tool[7]. To examine this, the capacity for transgenesis, genome editing, transient transgene
183 expression and the presence, integrity, and expression levels of RNAi-associated genes were analysed in
184 LAB and QLD (Figure S7). In both accessions, principal viral defence RNAi genes[37], *DCL2*, *RDR6*, *DRB4* and
185 *AGO2* have one expressed homeolog, both functional *DCL4* homeologs, and four expressed copies of
186 *AGO1*. The number, integrity and expression of these genes does not differ significantly between the
187 accessions, nor those of RNAi genes involved in chromatin remodelling or endogenous small RNA
188 production (Figure S7). *NbRDR1* is the exception. In LAB, there is a 72nt insertion that creates stop codons
189 towards the middle of the gene[38]. Curiously, the mRNA is full length and accumulates like that of its
190 uninterrupted QLD counterpart. Nonetheless, the truncated NbRDR1 protein in LAB is not acting as a
191 dominant negative as engineering early stop codons into the gene did not relieve the viral susceptibility
192 (Figure S8). To test whether the difference in RDR1 function might make QLD a superior or inferior
193 research tool and bio-platform to LAB, the accessions were assessed for ease and efficiency of
194 transformation and gene editing and level of transient gene expression from syringe and vacuum
195 infiltration (Figure E6,7; Table S13 Figure 4). In almost all of these respects they performed similarly.
196 However, LAB yielded a much higher level of transiently expressed antibody from vacuum agro-infiltration
197 (Figure 4B,C), is physically easier to patch-infiltrate, and has a faster generation time[14].

198

199 **Expansion and contraction of Transposable Elements**

200 Polyploidization is often accompanied by bursts of transposable element (TE) activity[39–42] and TEs,
201 especially the type 1 long terminal repeat class (LTRs) such as Gypsy metavirus (Gypsy), are highly
202 abundant in *Nicotiana*[34]. While Gypsy proliferation is obvious in the *N. benthamiana* genome, its content
203 (~1.5Gb) is more similar in size to those of the diploid *Nicotiana* species than to the allotetraploid *N.*
204 *tabacum* or the combined sum of the extant ancestral parental diploid relatives, *N. glauca* and *N. sylvestris*
205 (Figure 5A). A similar expansion of Gypsy content is evident in the recently reported pepper genome and

is one of the main causes for its increased size[43]. However, as a percentage of genome size, all of these *Nicotianas*, including *N. benthamiana*, are about 50% Gypsy or Gypsy-like sequence, suggesting that the decreased Gypsy content in *N. benthamiana* is due to whole chromosome loss rather than TE-mediated genome purging [44,45].

Unlike any other sequenced Solanaceous species genome, including the closely related diploid *N. attenuata* and the polyploid *N. tabacum*, the *N. benthamiana* genome shows evidence of dramatic, recent Copia element proliferation (Figure 5A, B). Examining in more detail four different loci in the subgenomes of LAB and QLD and comparing them with their counterparts in tomato and other *Nicotianas* (Figures E8-10) revealed a common theme of expansion of intergenic regions in the *Nicotianas* compared to tomato, which, as in pepper, is largely due to Gypsy elements which are now highly fragmented. A second theme is tandem duplication in *Nicotiana*, followed by extensive pseudogenization specifically in *N. benthamiana*. An abundance of recent, intact Copia elements is also evident in *N. benthamiana.* Insertion dating (Figure 5B) reveals that sustained periods of Copia mobility started around 2 MYA, reaching a peak around 750 KY, and are still occurring. This coincides with the divergence of LAB and QLD, dated at ~ 800 KYA[14], and recently inserted Copia elements are evident in close proximity to key genes in all four loci that we examined (Figures E8-10) suggesting that the recent mobility has played a major role in the genome's advancing diploidization and diversity. It is possible that the Copia explosion is common to all of the Australasian *Nicotianas* and, in conjunction with their allopolyploidy, this has possibly fueled the adaptation enabling the widespread success of the *Suaveolentes* across some of the harshest climatic and ecological regions in Australia.

Epigenetic landscape and sites of transgene integration

The epigenetic landscape of the LAB genome was examined for histone H3 methylation and acetylation, and cytosine methylation (Figures 5C-D, S9)[46]. Chromosomes 1,2,3,4,5, and to a lesser extent, 11 and 12, have a pronounced gradient of gene density across each chromosome which helps to reveal the correlation of high gene density with high levels of active histone marks (H3K4me3, H3K27ac). An inverse correlation of high gene density with repressive histone and DNA marks (H3K9me2 and CG and CHG methylation) is also apparent. These epigenetically repressed regions contain high levels of fragmented Gypsy elements whereas the active regions correlate with increased levels of intact Copia elements. The associations are also visible in the other chromosomes at a more localised level. The remarkably high level

of recent Copia element insertions into regions with high gene density and active histone marks also correlates with high levels of CHH methylation which are likely driven by active transcription of these TEs.

To investigate whether epigenetic landscape has an influence on transgene insertion in the *N. benthamiana* genome, stable transgenic lines and leaf patches agro-infiltrated with transgene-encoding constructs were analysed for their insertion locations. From 40 independent transgenic lines, 23 sites could be mapped, and whole genome sequencing of the infiltrated patches identified 144 integration sites (Figure 5D). When adjusted for chromosome size, there was no significant bias for integration into any specific chromosome (p=0.19). However, integration into the gene body and promoter elements was more frequent than random (Figure S10) and those inserting into intergenic regions were significantly closer to the gene borders (Figure S11). Transgene insertion into the gene body was at a much higher rate in transiently agroinfiltrated tissue than in stable transgenic lines, presumably because insertion-mediated dysfunctionality of some genes prevents whole plant regeneration but are not lethal in confined patches of somatic tissue. The average intergenic size for *N. benthamiana* is ~60Kb (Figure S12) and the majority of transgenes have been inserted within the 10Kb region adjacent to a gene. A similar bias is apparent for active copies of both Copia and Gypsy (Figure 5D, Figure S11). Coupled with the histone and cytosine methylation status data, this supports the notion that trangenes and TEs are more able to integrate into the open chromatin of genes and adjacent regions than into the condensed core of intergenic zones.

Diploidization and pathway dysfunction in *N. benthamiana*

The loss of 5 chromosomes from the ancestral allotetraploid with retention of ~50% of the genes in the genome as singletons (LAB sgA:10,075 sgB:11906; QLD sgA: 11,416 sgB:12,905) rather than homeologous pairs (Figure 2D, Tables S11A-H), indicates a loss of ~20K genes/genome over 5MY. This complies with the estimation that the ancestral allotetraploid genome had ~70K genes[31,32] and, coupled with LAB's genetic dysfunctions, explains the simple 3:1 Mendelian inheritance ratios of many traits in LAB x QLD crosses, such as virus susceptibility[14], nor-nicotine production and anthocyanin competence. In each of these, LAB has dysfunctional genes and pathways compared to QLD. The anthocyanin-regulating transcription factor (TF) locus shows tandem gene duplication with progressive gene dysfunction (Figure E8B). Even more striking diploidization is apparent in the nicotine synthesis regulating *ERF IX* TF locus (Figure E8A), the *RPM1*-like bacterial defence gene locus (Figure E9A) and the terpene biosynthesis *CYP736A* gene locus (Figure E9B). In all of these, there is evidence of recently inserted Copia elements, suggestive of their role

in the process. Diploid *Solanum* genomes and many non-solanaceous species exhibit high gene density bias towards the chromosome termini (Figure 5E-F). Interestingly, *N. benthamiana* chromosomes, especially 5-10 and 15-19, have a more uniform density. This unusual arrangement was likely caused by their formation through abundant inter-chromosomal recombination and by gene density dilution through the favoured insertion of TEs into the active chromatin of gene-rich regions.

Discussion

The exponential adoption of *Nicotiana benthamiana* as a model plant over the last two decades has produced vast amounts of data describing its responses to a wide spectrum of biotic and abiotic challenges, and this seems likely to continue unabated. Its use as a bioplatform to produce therapeutics has a similar trajectory. This dual role as a model species and non-food bioproduction platform, on top of the unmatched capacity for fast transient transgene analysis, has made *N. benthamiana* the chassis of choice for testing and implementing the most advanced engineering approaches in plant synthetic biology[47–49]. We have produced a high-quality genome assembly of the LAB strain of *N. benthamiana* with fully annotated gene models, microRNA families, transposable elements, epigenetic landscapes, and chromosomal subgenomic membership, and made this publicly available on an interactive web-based genome browser. This enables decades of previously obtained data to be placed into a broader context, provides an important aid for future research and biotechnology, and facilitates the involvement of the scientific community to expand and refine the resource. The high-quality genome assembly of QLD with its additional pathways and ~ 3,000 genes, and the details about genomic diversity of an additional 4 wild and 2 laboratory isolates, provide resources to greatly enhance metabolic, developmental, and evolutionary studies. This is not only relevant to *N. benthamiana* but also across the Solanaceae, as it brings the genome of a *Nicotiana* species to the same chromosomal level of completeness (>95%) as tomato, eggplant, potato, and pepper.

Compared to QLD, LAB is defective in many pathways including viral defence due to a dysfunctional RNA polymerase gene (*RDR1*), but both accessions have similar levels of expression and homoeolog retention for the other RNAi pathway genes. Although QLD has a greater genetic spectrum for metabolic and biotechnological engineering than LAB and similarly high transformation and gene editing efficiencies, its slower growth rate and lower yields of transiently expressed antibodies following vacuum agro-infiltration make LAB the preferred choice as a biofactory and research tool. However, QLD and LAB are highly interfertile (Figure S13) making them a powerful partnership for a wide range of molecular genetic and

comparative genomics approaches such as recombinant inbred (RIL) and epigenetic recombinant inbred (epiRIL) populations reminiscent of well-established model plant systems such as Arabidopsis, Maize and Rice.

N. benthamiana shows a recent explosion of Copia mobility and rapidly advancing diploidization. These two phenomena may or may not have a cause-effect relationship, but are apparently unique to this species, among sequenced *Nicotianas*, making it an excellent model species to study the course of diploidization and the dynamic balance of two subgenomes undergoing this process.

Acknowledgments

We thank the QUT High-Performance Computing team for their assistance in genome assembly and hosting the genome browsers. Work was supported by Australian Research Council Federation Fellowship, Laureate Fellowship, Discovery, and Centre of Excellence Awards (FF0776510 PW, FL160100155 PW, DP170103960 PW, and CE200100015 PW) and the European Commission Horizon 2020 program, project "Developing Multipurpose Nicotiana Crops for Molecular Farming using New Plant Breeding Techniques (NEWCOTIANA)", Grant Agreement n. 760331 DO, AB, GG, PW and 101094738 GG.

Author Contributions

BR, JA, ML, HJ, KD, JB, DO, GG, AB, CW and PW conceived and designed the project. Genome assembly and annotation was conducted by JA, ML, HJ, BR, VL, GA and VG-C. HiC data collected by PD, SZ, SBDC and JB. Gene editing by SR, FN and SH. GFP, anthocyanin, volatiles and antibody expression performed by LA, ZA, SR, BR, FF and ZL. Metabolic analysis by MS, BR, and GG. Chromosome allocation to subgenomes, and synteny analysis by JA, TC, ML and PW. PW, BR, JA and GG wrote the first drafts of the manuscript and all authors reviewed and edited the manuscript and approved the final version.

Competing Interest

The authors declare no competing interests.

Table 1: Genome assembly metrics of LAB and QLD compared with reference genomes. Various genome assembly quality criteria (L50, N50, BUSCO score) are used to compare *N. benthamiana* with the other available genomes. The values in brackets below *N. tabacum* and *N. attenuata* are those obtained from scaffold data alone.

Species/Accession	Scaffolds >500nt	Chr	L50	N50 (Mb)	Assembled Genome Length (Gb)	BUSCO% Complete v10
N. benthamiana LAB	19	19	10	145	2.75	C:98.1%[S:46.0%,D:52.1%]
N.benthamiana QLD	19	19	10	141	2.72	C:98.0%[S:47.5%,D50.5%]
Arabidopsis thaliana	5	5	3	23	0.12	C:99.2%[S:98.7%,D:O.5%]
Potato (dihaploid)	12	12	6	59	0.74	C:98.4%[S:96.6%,D:1.8%]
Tomato	12	12	6	61	0.72	C:97.8[S:96.8%,D:1.0%]
Eggplant	12	12	5	76	0.83	C:84.2%[S:82.7%,D1.5%]
Tobacco Chromosomes (Scaffolds)	24 (942,183)	24	9 (3,998)	84 (0.22)	1.74 (4.01)	C:82.6%[S:61.2%,D:21.4%] (C:96.8%[S:24.3%, D:72.5%1])
Capsicum	12	12	6.00	221	2.56	C:74.8%[S:73.7%,D:1.1%]
N. attenuata	12 (37,194)	12	498 (1,627)	66 (0.45)	0.73 (2.09)	C:48.5% [S:47.4%, D:1.1%] (C:98.1% [S:95.9%, D:2.2%])
Petunia axilaris	17,630	12	17,630	1.24	1.20	C:98.2%[S:95.6%,D:2.6%]
N. benthamiana LAB (USA vI.0.I)	52,890	19	1,718	0.44	2.49	C:98.2%[S:45.8%,D:52.4%]
Petunia inflata	35,907	12	35,907	0.88	1.17	C:97.9%[S:91.6%,D:6.3%]
N. benthamiana LAB (AU v0.5)	77,255	19	1,903	0.39	2.49	C:97.6%[S:47.5%,D:50.1%]
N. sylvestris	125,957	12	7,255	0.08	2.01	C:95.1%[S:93.3%,D:1.8%]
N. tomentosiformis	90,682	12	5,563	0.15	1.62	C:94.4%[S:92.6%,D:1.8%]
N. obtusifolia	20,758	12	2,189	0.05	3.50	C:94.3%[S:92.3%,D:2.0%]
N. otophora	420,947	12	14,141	0.03	2.32	C:76.0%D[S:74.3%,D:1.7%]

Figure 1. Phenotypic and biochemical diversity of *N. benthamiana*. **A)** Proposed phylogeny and origin of the *Suaveolentes* section compared to other Nicotianas. Chromosome numbers are indicated for each *Suaveolentes* species. Species highlighted by an asterisk are extant relatives of the putative parents of *N. benthamiana* and *N. tabacum*. **B)** Distribution of *N. benthamiana* in Australia (chequered regions). The physical locations of isolated *N. benthamiana* accessions reported in this study are shown by pins, and traditional indigenous trading routes are shown by red lines. **C)** Profiles of average emission of selected floral volatile compounds from LAB and QLD over a 24-hr period. Dark blue: Benzyl alcohol. For other compounds see Figure E1. Data are presented as mean values +/- standard error (n=4 per sample point). **D)** Anthocyanin production 5 days after transient expression of AN-like MYB in LAB and QLD; right panels show protoplasts isolated from LAB and QLD infiltrated patches (n=5). The scale bar is 50 um. **E)** Comparison of the

accumulation of nicotine and nornicotine in flowers and leaves of LAB and QLD. The biochemical conversion of nicotine to nornicotine, mediated by the CYP82E demethylase (see Figure E9), is shown on the right. Data are presented as mean values +/- standard error (n=4). **F)** Comparison of the accumulation of HGL-DTGs in flowers and leaves of LAB and QLD. The schematic biochemical pathway is shown on the right. Data are presented as mean values +/- standard deviation (n=4). Comparison of the accumulation of HGL-DTGs in flowers and leaves of LAB and QLD. The biochemical pathway is shown on the right. Data are presented as mean values +/- standard error (n=2 biologically independent samples).

Figure 2. Subgenome and homeolog organization in *N. benthamiana.* A) The left hand Circos plot depicts the locations of the syntenic blocks (1Mbp) of *N. tomentosiformis* (blue) and *N. sylvestris* (red) on the *N. tabacum* genome, highlighting the subgenomes and their respective contribution to the subgenome structure of this species. The right hand Circos plot similarly locates the syntenic blocks of *N. tomentosiformis* (blue), *N. sylvestris* (red) and *N. glauca* (purple) on the *N. benthamiana* LAB genome, highlighting the difficulty in assigning ancestry for subgenomes in this species, which is characterised by extensive rearrangement of blocks between individual chromosomes. The lines in the centre join syntenic regions, highlighting the fragmentation of the *N. benthamiana* genome. **B)** Dot plot showing the relationship between the LAB and QLD chromosomes (central continuous line in the far-left panel) and the fragmented syntenic relationship between the subgenomes. Comparison of the *N. tabacum* genome consisting of 2 subgenomes with clear relationships to *N. sylvestris* and *N. tomentosiform*is revealed a fragmented relationship with *N. benthamiana* chromosomes. **C)** Dendrogram highlighting the chromosome pairs and the 3 orphan chromosomes (annotated to 9, 10 and 19). **D)** Retention and relocation of homeologous genes in *N. benthamiana* LAB and QLD genomes. Percentage values outside and within brackets are those for LAB and QLD, respectively, and show that about half of the original homeologous pairs have lost one member.

Figure 3. Gene block conservation across the Solanaceae and segmental allopolyploidisation in *N. benthamiana.* A) Waterfall plot showing the syntenic relationships between LAB, QLD and

other related species as generated by SynVisio. **B)** Fraction of orthologous gene clusters in different Solanaceae chromosomes, highlighting a high conservation of chromosomes 1-4, and a declining conservation of remaining chromosomes; chromosome numbering largely follows the tomato-potato system. **C)** A Gibson Venn diagram showing the number of gene family clusters that are shared amongst LAB, *N. sylvestris* and *N. glauca.* **D)** Overlay of *N. glauca* (blue bars within chromosomes) and *N. sylvestris* (red) orthologous genes on LAB chromosomes; Grey/blue lines connecting chromosomes link syntenic blocks among the matched subgenome chromosomes. **E)** Circos plot of the physical distribution of syntenic blocks of tomato chromosomes 9-12 overlaid onto the LAB genome (track a), showing extensive fragmentation across the remaining LAB chromosomes. In contrast, an overlay of the syntenic blocks of Tomato chromosomes 1-4 onto the LAB genome clearly demonstrates the conservation of both sequence and location (track b). Track c shows the gene density across the LAB chromosomes.

Figure 4. Comparison of transient expression in LAB and QLD of green fluorescent protein (GFP) by syringe agroinfiltration and antibody production by vacuum agroinfiltration. A) Transient expression of GFP in LAB and QLD. Quantitative polymerase chain reaction (qPCR) cycle threshold (Ct) values were measured and ΔCt calculated as the difference in Ct between the gene of interest (GFP) and the reference gene (GAPDH) for each sample. GFP expression levels are represented underneath each leaf as ΔCt. All reactions were performed in triplicate for each cDNA sample. All experiments were performed in eight independent biological replicates. The average ΔCt of LAB and QLD was 4.8 and 4.7, respectively. Statistical analysis of the two-tailed student's t-test (P=0.7972) and z-test (P=0.9949) shows that there was no significant difference between GFP expression levels in the two ecotypes. Scale bar is 1 cm. B) Antibody concentration in total soluble protein extracts from Lab (white) and QLD ecotypes (grey) measured by protein A bio-layer interferometry in µg/mg of tissue fresh weight (FW). P-values determined by Mann-Whitney U test comparing between ecotypes. For 'n', samples are biologically independent transient infiltrations, sampled at 7 days post infiltration. Box and whisker plot interpretation: each box spans the interquartile range with the ends of the box being the upper and lower quartiles. The median is represented by a vertical line inside the box. Whiskers outside the box

extend to the highest and lowest observations. C. SDS-PAGE showing protein A purified Trastuzumab under reducing condition, similar results were observed in 3 independent replicates (n=3).

Figure 5. Transposon, epigenetic landscapes, and gene density of *N. benthamiana*. A) Relative complements of transposon and non-transposon content in *A. thaliana*, *V. vinifera* and key Solanaceous and *Nicotiana* species is presented as their calculated genome content in Gb. The dashed box for *N. glauca* indicates the genome size calculated from k-mer analysis (4.5Gb) while the composition of the genome is based on the current assembly of 3.2-3.5Gb. N.B. Many Gypsy-like sequences are present in the "Other TE" category in *N.benthamiana*. **B)** Estimated dates of LTR-retrotransposon insertion, calculated by sequence comparison between the LTRs of individual element insertions, in *N. benthamiana* LAB and QLD, compared to *N. attenuata* and *N. tabacum*. A clear and ongoing large burst of Copia element activity is evident in both LAB and QLD, which is absent in both *N. attenuata* and *N. tabacum*. The reported burst of Gypsy activity in *Nicotiana* species appears to predate the 6 MYA limit of our analysis. **C)** A Circos plot depicting the chromatin landscape as compared to gene content in LAB. Tracks a and b represent respectively the location of permissive histone marks H3K27ac and H3K4me3 within each LAB chr. Track c depicts the gene density across the LAB genome, while tracks d and e represent the location or repressive histone marks H3K9me2 and H3K27me3, respectively. **D)** Circos plot depicting the comparative locations of transgene insertions, LTR-retrotransposon insertion and methylation marks across LAB chromosomes. Track a: Transgene insertion sites; red 'ticks' represent insertions derived from stable transformation, blue 'ticks' represent insertions derived from transient agroinfiltration. Track b: Insertions of intact Copia TEs (i.e. containing matching LTRs and complete internal sequences). Track c: Insertion of all annotated Copia TEs, including fragmented elements. Track d: Distribution of CHH methylation marks. Track e: Gene density across the LAB genome. Track f: Insertions of all annotated Gypsy TEs, including fragmented elements. Track g: Distribution of CG methylation marks. Track h: distribution of CHG methylation marks. The innermost circle represents the numbered chromosomes. **E)** Distribution of gene densities on the chromosomes of potato and tomato. **F**) Distribution of gene densities on the

chromosomes of LAB and QLD genomes.

References

1. Knapp, S., Bohs, L., Nee, M. & Spooner, D. M. Solanaceae—A Model for Linking Genomics with Biodiversity. *Comparative and Functional Genomics* vol. 5 285–291 Preprint at https://doi.org/10.1002/cfg.393 (2004).
2. Narayan, R. K. J. Nuclear DNA changes, genome differentiation and evolution inNicotiana (Solanaceae). *Plant Syst. Evol.* **157**, 161–180 (1987).
3. Clarkson, J. J., Kelly, L. J., Leitch, A. R., Knapp, S. & Chase, M. W. Nuclear glutamine synthetase evolution in Nicotiana: phylogenetics and the origins of allotetraploid and homoploid (diploid) hybrids. *Mol. Phylogenet. Evol.* **55**, 99–112 (2010).
4. Marks, C. E., Ladiges, P. Y. & Newbigin, E. Karyotypic variation in Nicotiana section Suaveolentes. *Genet. Resour. Crop Evol.* **58**, 797–803 (2011).
5. Bally, J. *et al.* Nicotiana paulineana, a new Australian species in Nicotiana section Suaveolentes. *Aust. Syst. Bot.* **34**, 477–484 (2021).
6. Byrne, M. *et al.* Birth of a biome: insights into the assembly and maintenance of the Australian arid zone biota. *Mol. Ecol.* **17**, 4398–4417 (2008).
7. Bally, J. *et al.* The Rise and Rise of Nicotiana benthamiana: A Plant for All Reasons. *Annu. Rev. Phytopathol.* **56**, 405–426 (2018).
8. Schiavinato, M., Marcet-Houben, M., Dohm, J. C., Gabaldón, T. & Himmelbauer, H. Parental origin of the allotetraploid tobacco Nicotiana benthamiana. *Plant J.* **102**, 541–554 (2020).
9. Schiavinato, M., Bodrug-Schepers, A., Dohm, J. C. & Himmelbauer, H. Subgenome evolution in allotetraploid plants. *Plant J.* **106**, 672–688 (2021).

10. Khafizova, G., Dobrynin, P., Polev, D. & Matveeva, T. Whole-genome sequencing of Nicotiana glauca. *bioRxiv* 211482 (2017) doi:10.1101/211482.
11. Usade, B. *et al.* The genome and metabolome of the tobacco tree, Nicotiana glauca: a potential renewable feedstock for the bioeconomy. *bioRxiv* 351429 (2018) doi:10.1101/351429.
12. LeBlanc, Z., Waterhouse, P. & Bally, J. Plant-Based Vaccines: The Way Ahead? *Viruses* **13**, (2020).
13. Waterhouse, P. M. & Helliwell, C. A. Exploring plant genomes by RNA-induced gene silencing. *Nat. Rev. Genet.* **4**, 29–38 (2003).
14. Bally, J. *et al.* The extremophile Nicotiana benthamiana has traded viral defence for early vigour. *Nat Plants* **1**, 15165 (2015).
15. Drapal, M., Enfissi, E. M. A. & Fraser, P. D. Metabolic changes in leaves of N. tabacum and N. benthamiana during plant development. *J. Plant Physiol.* **265**, 153486 (2021).
16. Drapal, M., Enfissi, E. M. A. & Fraser, P. D. Metabolic effects of agro-infiltration on N. benthamiana accessions. *Transgenic Res.* **30**, 303–315 (2021).
17. Steppuhn, A., Gase, K., Krock, B., Halitschke, R. & Baldwin, I. T. Nicotine's defensive function in nature. *PLoS Biol.* **2**, E217 (2004).
18. de Boer, G. & Hanson, F. E. Feeding responses to solanaceous allelochemicals by larvae of the tobacco hornworm, Manduca sexta. *Entomol. Exp. Appl.* **45**, 123–131 (1987).
19. Snook, M. E. *et al.* Hydroxygeranyllinalool Glycosides from Tobacco Exhibit Antibiosis Activity in the Tobacco Budworm [Heliothis virescens (F.)]. *J. Agric. Food Chem.* **45**, 2299–2308 (1997).
20. Jassbi, A. R., Zamanizadehnajari, S., Kessler, D. & Baldwin, I. T. A New Acyclic Diterpene Glycoside from Nicotiana attenuata with a Mild Deterrent Effect on Feeding Manduca sexta Larvae. *Zeitschrift für Naturforschung B* **61**, 1138–1142 (2006).
21. SGN EST search result - sol genomics network. https://solgenomics.net/search/est.pl?request_id=SGN-

E1214852&request_from=0&request_type=automatic&search=Search.

22. Nakasugi, K., Crowhurst, R., Bally, J. & Waterhouse, P. Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana. *PLoS One* **9**, e91776 (2014).

23. Ruiz, M. T., Voinnet, O. & Baulcombe, D. C. Initiation and maintenance of virus-induced gene silencing. *Plant Cell* **10**, 937–946 (1998).

24. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).

25. Howe, K. *et al.* Significantly improving the quality of genome assemblies through curation. *Gigascience* **10**, (2021).

26. Lawniczak, M. K. N. *et al.* Standards recommendations for the Earth BioGenome Project. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).

27. Philips, J. G. *et al.* The widely used Nicotiana benthamiana 16c line has an unusual T-DNA integration pattern including a transposon sequence. *PLoS One* **12**, e0171311 (2017).

28. Lorenc, M. T. *et al.* Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology* **1**, 370–382 (2012).

29. Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5135–5140 (2014).

30. Dunn, N. A. *et al.* Apollo: Democratizing genome annotation. *PLoS Comput. Biol.* **15**, e1006790 (2019).

31. Barchi, L. *et al.* A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Sci. Rep.* **9**, 11769 (2019).

32. 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**, 448 (2017).

33. Brockmöller, T. *et al.* Nicotiana attenuata Data Hub (NaDH): an integrative platform for exploring genomic, transcriptomic and metabolomic data in wild tobacco. *BMC Genomics* **18**, 79 (2017).
34. Xu, S. *et al.* Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 6133–6138 (2017).
35. Bandi, V. & Gutwin, C. Interactive Exploration of Genomic Conservation. *Graphics Interface* Preprint at https://doi.org/10.20380/GI2020.09 (2020).
36. Gaeta, R. T. & Chris Pires, J. Homoeologous recombination in allopolyploids: the polyploid ratchet. *New Phytol.* **186**, 18–28 (2010).
37. Qin, C. *et al.* Roles of Dicer-Like Proteins 2 and 4 in Intra- and Intercellular Antiviral Silencing. *Plant Physiol.* **174**, 1067–1081 (2017).
38. Yang, S.-J., Carter, S. A., Cole, A. B., Cheng, N.-H. & Nelson, R. S. A natural variant of a host RNA-dependent RNA polymerase is associated with increased susceptibility to viruses by Nicotiana benthamiana. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6297–6302 (2004).
39. Grandbastien, M.-A. *et al.* Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet. Genome Res.* **110**, 229–241 (2005).
40. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* **46**, 270–278 (2014).
41. Kuang, H. *et al.* Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res.* **19**, 42–56 (2009).
42. Naito, K. *et al.* Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**, 1130–1134 (2009).
43. Liao, Y. *et al.* The 3D architecture of the pepper genome and its relationship to function and evolution. *Nat. Commun.* **13**, 3479 (2022).

44. Lee, S.-I. & Kim, N.-S. Transposable elements and genome size variations in plants. *Genomics Inform.* **12**, 87–97 (2014).
45. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* **12**, 1075–1079 (2002).
46. An, J. *et al.* J-Circos: an interactive Circos plotter. *Bioinformatics* **31**, 1463–1465 (2015).
47. Mitiouchkina, T. *et al.* Plants with genetically encoded autoluminescence. *Nat. Biotechnol.* **38**, 944–946 (2020).
48. Brophy, J. A. N. *et al.* Synthetic genetic circuits as a means of reprogramming plant roots. *Science* **377**, 747–751 (2022).
49. Bernabé-Orts, J. M. *et al.* A memory switch for plant synthetic biology based on the phage φC31 integration system. *Nucleic Acids Res.* **48**, 3379–3394 (2020).

Online Methods

Plant lines

N.benthamiana LAB, NT, SA, WA, QLD, and NWA accessions have been described [1]. The EuLAB isolate is the extensively used GFP-expressing transgenic line (16c) produced in David Baulcombe's laboratory, Sainsbury Institute, UK[2,3] and USA-LAB have been described [4]**.** Plants were grown in a custom soil mix (UQ23 supplemented with Osmocote® slow release fertiliser) under controlled environmental conditions at a constant temperature of 25 °C with a 16-hour photoperiod and 8-hour dark period.

RNA-Seq

Total RNA was isolated from four tissues (leaf, flower, stem, root) and seedlings (10 days) of LAB (6 weeks) and QLD (7 weeks) at the same developmental stage using TRIzol™ Reagent according to the manufacturer's instructions. Libraries were constructed in triplicate for each tissue using NEBNext® ultra™ RNA Library Prep Kit for Illumina®, size selected (average 300nt), and sequenced on an Illumina HiSeq 2000/2500 system to produce 150 bp paired-end reads.

Extraction and analysis of secondary metabolites from plant tissues

Flower, leaf, stem, and roots were sampled as described for RNAseq and two biological replicates (individual plants) of the same samples of LAB and QLD were used for the metabolic analysis. Tissues were freeze-dried and homogeneously grounded in liquid nitrogen.

The semi-polar fraction was extracted from lyophilized ground tissue (3 mg for flower and root, and 5 mg for leaf and stem tissues) with 75% Methanol/0.1% v/v Formic acid, spiked with 0.25 µg/mL formononetin (Sigma-Aldrich) as internal standard. Metabolites were extracted at room temperature by continuous agitation for 30 min in MM 400 at 20 Hz. Samples were centrifuged at 20,000g for 20 min, and 0.6 ml of the supernatant was transferred into filter (PTFE) vials for LC/MS analysis (0.2 µm pore size). Two independent extractions and analyses were performed for each biological replicate. LC conditions have been already described[5]. Five µL of the filtered extract was injected into the LC/HESI/MS system, using a Q-exactive mass spectrometer (ThermoFisher Scientific). The ionization was performed using the heated electrospray ionization (HESI) source, with nitrogen used as sheath and auxiliary gas, and set to 35 and 10 units, respectively. The capillary temperature was 250°C, the spray voltage was set to 3.5 kV, the probe heater temperature was 330°C, and the S-lens RF level was set at 50. The acquisition was performed with the FTMS mass range of 110-1,600 m/z both in positive and negative ion mode, with the following parameters: resolution 70,000, microscan 1, AGC target 1e6, and maximum injection time 100. Dd-MS2 parameters were as follows: resolution 17,500, intensity threshold 4.0e4, AGC target 2e4, maximum IT 50 ms, TopN 5, stepped NCE 15, 25, 40. All the chemicals and solvents used during the entire procedure were of LC/MS grade (Chromasolv, Merck Millipore).

Metabolic diversity was evaluated by comparing the MS spectra (positive ion mode) using SIEVE software (Thermofisher Scientific)[5]. The LC/MS spectra were processed by comparing tissues from each ecotype, only metabolites accumulating to levels of > 2-fold change and pval<0.05 (t-test) between the two ecotypes were selected. Metabolites were identified based on accurate masses in full MS together with MS2 spectra and/or authentic standards, using KEGG (https://www.genome.jp/kegg/compound/), Metfrag (https://ipb-halle.github.io/MetFrag/projects/metfragweb/) and PubChem mass databases (ST3) (https://pubchem.ncbi.nlm.nih.gov/). Relative levels of accumulation of investigated metabolites were measured and normalized relative to DW and the internal standard, to correct for extraction and injection variability, as described[5].

Whole Plant Vacuum Infiltration and Antibody Purification

Small-scale trastuzumab expression studies were performed using 5 to 6-week-old *N. benthamiana* plants. *Agrobacterium tumefaciens* strain GV3101 containing plasmids with expression cassettes for trastuzumab light chain (LC), trastuzumab heavy chain (HC), P19 and galactosyl transferase (https://www.plantformcorp.com/) were centrifuged at 12000 g for 30 min then resuspended in infiltration buffer to an OD600=0.2. The infiltration solution was poured into 2 L beakers, filling each beaker to the rim. The aerial portions of N. benthamiana plants were submerged into the infiltration solution and placed in a 15-gallon vacuum chamber (Best Value Vacs Cat #BVV15G). Using a vacuum line, a vacuum was applied until the pressure on the chamber reached -25 in. -Hg, then held for 3 min and slowly released. *N. benthamiana* plants were then removed from solution and returned to the growth chamber. Leaf tissue was harvested seven days post infiltration and stored at -80oC until processing.

Frozen infiltrated plant tissue was homogenized in liquid N2 with a mortar and pestle then combined with 3 volumes of 4oC PBS buffer pH 7.4. The homogenate was then centrifuged at 16000 g for 30 min at 4oC. The TSP was then passed through a 0.45 µm filter into a clean tube. The antibody was then purified according to the manufacturer's instructions supplied with the Protein G HP SpinTrap kit (GE HealthCare Cat# 28903134) using the standard purification protocol.

Whole Genome Sequencing

High molecular weight genomic DNA from leaves or leaf nuclei of *N. benthamiana* LAB and QLD ecotypes were extracted as described [6] and used for whole genome sequencing (Illumina, PacBio and Oxford Nanopore; see Figure S3). Illumina and PacBio sequencing was conducted by the Central Analytical Research Facility (CARF), Queensland University of Technology (QUT-CARF) and Nanopore sequencing by the Australian Genome Research Facility (AGRF), Melbourne. The quality of the assemblies was determined using Merqury software (version 1.3)[7]. LAI scores were determined using the annotation obtained from the EDTA TE annotation pipeline[8] and using the LAI sub-package of the LTR-retriever[9] package according to Ou et al. 2018[10] (see also https://github.com/oushujun/EDTA/wiki/Calculate-LAI-from-EDTA-GFF3-files).

Genome assembly

609 The assembly pipeline is summarized in Figure S3. LAB and QLD contigs were assembled using CANU
610 (version 1.81)[11] and SparseAssembler k-mer 77 (version 20160205)/DBG2OLC (version 20160205)/Racon
611 (version 1.3.2)[12–14], respectively. Bionano optical mapping[15] gave 44 and 37 super scaffolds for LAB and
612 QLD, respectively, with contiguity statistic N50 values of 122 and 130Mbp. Juicer (version 1.6)[16] and 3D-
613 DNA (branch 201008)[17] were used to generate Hi-C data and pre-assembly files. HiC libraries were
614 produced as described by Dong et al.[18], sequenced using the Illumina platform, and the aligned fragments
615 from Juicer were further refined using Juicebox (version 2.12)[19] and Citrus
616 (https://github.com/anjiyuan/Citrus) to produce chromosome level assemblies. LR_Gapcloser[20] (version
617 1.1) was used to close gaps with long reads to complete our genome assemblies. Afterwards, both
618 assemblies were polished with Illumina reads using Pilon[21] (version 1.23). Finally, Mercury[7] (version 1.3)
619 was used to categorize assembly quality based on EBG[22]. First, k-mer for DNA Illumina sequence was
620 generated by running the tool with "meryl k=21 count output xxx.meryl xxx.fastq.gz" and then generating
621 k-mer completeness and QV value with "merqury.sh xxx.meryl <gene fasta> <prefix-output>". The
622 bioinformatic analyses were performed at the High Performance Computing (HPC) facility, QUT, and on
623 Flashlite on QRIScloud, Australia.

624

625 **Gene annotation**

626 HISAT2 (version 2.1.0)[23] generated BAM files using pooled RNAseq data (leaf, root, stem and seed) and
627 Scallop (version 0.10.5)[24] was used to identify transcripts from the pooled RNAseq data. Transdecoder
628 (https://github.com/TransDecoder/TransDecoder/) identified the coding and UTR regions. AUGUSTUS
629 (version 3.2.3)[25] was used to predict all possible transcripts based on the genome sequence. Combining
630 the two gene annotations[26], gave 267,000 and 255,000 genes for LAB and QLD, respectively. To filter out
631 low-confidence predicted genes, coding sequences of all the predicted genes were BLAST-searched[27]
632 against the NCBI NR (non-redundant) gene database and Solanaceae plants (tomato, potato, *N. attenuata*,
633 *N. tabacum*) with the "identity" parameter gradually reduced until the BUSCO (version 4.0.5)[28] score did
634 not increase. These were identity values of 86% (LAB) and 83% (QLD). To simplify the gene annotation,
635 only one isoform (the longest) was retained where there appeared to be overlapping genes.
636 Supplementing these high-confidence genes with those lost in the analysis but identified by Scallop gave
637 45,796 and 49,636 genes for LAB and QLD, respectively. Gene mapping was undertaken by BLAST

searching Tomato (https://solgenomics.net/ftp/tomato_genome/assembly/build_4.00/, v4.0), *N. attenuata* (https://www.ncbi.nlm.nih.gov/assembly/GCF_001879085.1/, including scaffolds) and *N. tabacum* (https://solgenomics.net/ftp/genomes/Nicotiana_tabacum/edwards_et_al_2017/) genomes with the sequences of gene coding regions from the LAB genome. Default BLAST settings were used.

Protein Cluster Analysis

Orthofinder version 2.5.4[29] (using default settings) identified orthologous relationships among LAB, QLD, identified *N. tabacum*, *N. sylvestris*, *N. tomentosiformis*, *N. glauca*, *A. thaliana, V. vinifera, S. lycopersicum, and S. tuberosum.* The UpSet plot in supplementary Figure 9 is generated using UpSetR package[30]. See Supplementary Table S7C for details about the genomes used.

TE annotation

The EDTA pipeline (version 2.0.0)[8] (https://github.com/oushujun/EDTA); last accessed 22/09/2022) was used to annotate the repeat element space for LAB, QLD, *N. attenuata,* and *N. tabacum* with the following initiating command:

>EDTA.pl --genome <genome fasta> --species others --step all -u --sensitive 0 --anno 1 --threads 48.

The annotation of the *N. tabacum* genome only made use of the chromosome assembly available from the Sol Genomics Network (https://solgenomics.net/organism/Nicotiana_tabacum/genome; file Nitab-v4.5_genome_Chr_Edwards2017.fasta.gz). The -u flag generates a file (*EDTA_raw/LTR/*.pass.list), containing estimations of LTR insertion times from LTR-retriever[9] a component part of the EDTA pipeline. The estimation of insertion time is based on the number of polymorphisms calculated between the LTR sequences of intact LTE-TEs. Due to the lack of an accurate estimation of the neutral mutation rate in *N. benthamiana*, the default rate was set to that calculated for rice: 1.3e-8 substitutions per base pair per year[8].

MicroRNA annotation

The mature microRNA sequences from 79 plant species (Table S8E) were retrieved from miRbase (release 21; https://www.mirbase.org/) and used to identify miRs in *N. benthamiana* using Bowtie (version 2.0)[31]. To avoid missing IsomiRs, possible mature miRNA sequences with one mismatch were also identified using miRPlant (version 6)[32]. The expression levels of each miR and its precursor transcript were calculated from pooled data of libraries of small RNA and RNAseq reads (from this and previous studies[33,34]).

SNP calling

All Illumina genomic paired-end reads from each ecotype were aligned to the LAB and QLD assemblies using bowtie2[35] (version 2.3.5). Duplicate reads were removed from each BAM file with Picard toolkit's (https://broadinstitute.github.io/picard/) (version 2.19), MarkDuplicates (*picard -Xmx25g MarkDuplicates ASSUME_SORT_ORDER=coordinate REMOVE_DUPLICATES=true*), and SAMtools[36] (version 1.10) was used to keep unique (*samtools view -Sb -q 40*) and proper pair-end reads (*samtools view -@ 1 -hb -f 0x2 -F 2316*). Each read ID in the BAM file was modified by adding the ecotype's id using generate_subset_BAM.py from the SGSautoSNP[37] pipeline (version 2.001). Next, BAM files for

each cultivar were merged using SAMtools to produce BAM files for LAB and QLD. Finally, The SGSautoSNP.py script was used with default parameters.

Chip-Seq

Cross-linking, chromatin isolation, nuclei lysis, chromatin shearing, and immunoprecipitation were carried out as described by Ranawaka *et al*.[6]. Antibodies against two active histone marks, anti-histone-H3-tri-methyl-K4 (Abcam ab8580) and anti-histone-H3-acetyl-K27 (Abcam ab4729), and two repressive histone marks, anti-histone-H3-tri-methyl-K27 (Abcam ab6002) and anti-histone-H3-di-methyl-K9 (Diagenode C15410060) were used in the immunoprecipitation step to generate the genome-wide histone modification landscapes of LAB and QLD. Libraries (two replicates per histone modification and control input) were prepared using NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (E7645S) as per the manufacturer's specifications. ChIP-seq libraries of H3K9me2 were sequenced at the Central Analytical Research Facility (CARF), Queensland University of Technology (QUT-CARF), using Illumina NextSeq® 500 with the output of 75 bp paired-end reads (TG NextSeq® 500/550 High Output Kit v2, 75 cycle, TG-160-2005). Libraries of H3K4me3, H3K27me3, and H3K27ac were sequenced at Novogene International Private Limited (Singapore) on the Illumina HiSeq® 2000/ 2500 system to produce 150 bp paired-end reads and analysed using the Galaxy platform (https://usegalaxy.org.au)[38]. Paired-end reads were aligned against LAB and QLD genome assemblies using bowtie2 (version 2.4.2) with default settings[31]. Alignments with MAPQ < 40 were discarded prior to downstream analyses to ensure homeolog specificity and accuracy. The deepTools, bamCompare[39], was used to quantify and visualise histone marks across genes.

Whole Genome Bisulfite Sequencing

Whole genome bisulfite sequencing samples were prepared with genomic DNA extracted from the same tissues used for ChIP-seq. Leaf genomic DNA from three replicates was extracted using a DNeasy Plant Mini Kit (QIAGEN, 69104). The bisulfite conversion of the DNA was carried out using the EZ DNA Methylation-Gold™ kit (ZYMO, D5005), and the bisulfite-treated DNA libraries were constructed using the Illumina TruSeq DNA sample prep kit, following the manufacturer's instructions. The library preparation and the subsequent next-generation sequencing were completed by Novogene HK Company Limited (Hong Kong Subsidiary). Paired-end read (150 bp) sequencing of the bisulfite-treated DNA libraries was performed using an Illumina HiSeqX system.

Methylation analysis

The high-quality reads from WGBS samples were aligned to LAB and QLD genome assemblies using the default settings of the Bismark program (Version 0.19.0)[40]. PCR duplicates were removed with the deduplicate_bismark implemented in the Bismark program (version 0.19.0). Reads were mapped to the non-methylated chloroplast genome as a control to calculate the sodium bisulfite conversion rate of unmethylated cytosines which was >99.9% for all replicates (three replicates from each LAB and QLD). The cytosine methylation level was calculated using the bismark_methylation_extractor in Bismark (version 0.19.0). The methylation ratio of cytosine was calculated as the number of methylated cytosines divided by the number of reads covering that position.

Calculation of relative expression levels of A and B subgenome homeologs

The MCScanX toolkit [41] was used to identify intraspecies syntenic blocks using protein sequences and chromosomal locations of genes (evalue 1e-10, max-target-seqs 6, masking 1, max-hsps 1). SynVisio[42], an interactive multiscale synteny visualisation tool for McScanX, was used to visualise the gene-level collinearity. Genes in syntenic blocks were identified as homeologs, and the genes that could not find their homoeologous partners were identified as singletons. The average transcripts per million (TPM) expression of genes in each tissue type was calculated (average expression per tissue). Then, using the average expression of each gene per tissue, the global expression across all tissues was calculated. Global expression > 0.5 TPM was used for downstream analysis. Values of this combined analysis were used to determine the relative expression of homeologs. The homoeologous pairs were defined as expressed when the sum of the 1 and 2 subgenome homeologs was > 0.5 TPM. This filtration included duplicate pairs where only a single homeolog was expressed. To standardise the relative expression of homeologs, the absolute TPM for each gene within the duplicate pair was normalised as follows. A and B represent the genes corresponding to the A and B homeologs in pairs.

Relative Expression of A=TPM(A)/(TPM(A)+TPM(B))

Relative Expression of B= TPM(B)/(TPM(A)+TPM(B))

The Kruskal-Wallis test was performed to statistically determine the homoeolog expression bias between subgenomes. Overrepresentation analysis was conducted using Fisher's Exact Test. All the genes in *N. benthamiana* were blasted, mapped, and annotated using the Blast2Go suite[43] and used as the background for the overrepresentation analysis. Highly suppressed genes in both subgenomes were assessed. Genes with a p-value < 0.05 were considered significantly overrepresented.

Identification and phylogenetic analysis of ERF189, NBS-LRR RPM1-like, Anthocyanin R2R3 Myb and Nicotine demethylase CYP82E genes

ERF189, *NBS-LRR RPM1*-like, Anthocyanin *R2R3 Myb* and *CYP82* genes in *N. benthamiana* were identified based on sequence homology using *N. attenuata* protein sequences (http://nadh.ice.mpg.de/NaDH/others/data) as query sequences for the tBLASTn function on Apollo (https://apollo.nbenth.com/). *N. attenuata* CYP82 (NiAv7g20333) was identified by sequence similarity to tobacco CYP82E4, a demonstrated nicotine demethylase gene[44]. Phylogenetic trees were built using the identified nucleotide sequences and their available counterparts in other *Nicotiana* sp. (*N. attenuata*, *N. tabacum*, *N. sylvestris*, *N. tomentosiformis*) aligned using Muscle (version 3.8)[45] . The best nucleotide substitution model was estimated based on jModeltest2 (version 2.1)[46] and a tree constructed for each gene family using MrBayes (version 3.2.6)[47] .

Transgene insertion analysis

Agrobacterium tumefaciens (GV3101) transformed with a 35s-GFP-OCS construct (pBEN0317) was infiltrated into 4- weeks- old *N. benthamiana* leaves. After five days, agroinfiltrated leaves were collected. Total genomic DNA was extracted using the ISOLATE II Plant DNA Kit Bioline (BIO-52070) and pooled before library preparation using TruSeq® DNA Library Prep Kits (FC-121-2001). Sequencing was performed using the Illumina® Hiseq 2000 platform. Paired-end reads were mapped to pBEN0317 binary vector using Burrows-Wheeler (BWA-MEM) (Version 0.7)[48]. To determine the T-DNA integration events, all split reads that partially overlapped the T-DNA region's left and right borders were extracted and searched using BLASTn against the *N. benthamiana* genome. Reads with a percentage identity higher than 85% and an E-value less than $1x10^{-5}$ were selected as high-confidence transgene integration sites. A different approach was used to identify the broken reads. Reads were initially mapped to the *N. benthamiana* genome and mapped reads whose mate is unmapped were extracted using Samtools view[36]. The filtered bam file was converted to fastq using bedtools Convert BAM to FastQ[49]. Reads were then BLASTed to the pBEN0317 vector. The reads which mapped to vectors with an E-value of less than $1x10^{-5}$ and more than a 100 bp alignment were then BLASTed to the *N. benthamiana* genome. Reads with high identity (>95%) and ~50% coverage were identified as integrated T-DNA into the plant genome. For the stable transformation analysis, leaf tissues were collected from 5 weeks old *N. benthamiana* stable transgenic independent lines generated using pFN117 (Cas9) and pUQC-GFP-(218). Genomic DNA was extracted following the Cetyltrimethylammonium bromide (CTAB) method. Nested, insertion-specific primers for the right

borders (RB1, RB2, and RB3) of pFN117 and pUQC-GFP-(218)-A were designed. Arbitrary degenerate primers (AD primers) and High-Throughput Thermal Asymmetric Interlaced Polymerase Chain Reaction (ht-TAIL-PCR) program were as described by Singer & Burke[50]. Purified PCR products were directly Sanger sequenced using RB3 primer, and the insertion sites were identified through a BLASTn search against the *N. benthamiana* genome. The number of stable and transient T-DNA insertion sites that intersect gene body, promoter, terminator, and transposable elements were determined using the bedtools Intersect tool (version 2.30.0)[49] and the length to the closest gene from the insertion site was calculated using RnaChipIntegrator (Version 1.1.0) (https://github.com/fls-bioinformatics-core/RnaChipIntegrator). The z-score test for two population proportions was used to determine the significant difference between 10kb, 10-20kb, 20-30kb and 30-40kb intervals from all stable, transient transgene insertion sites and randomly selected sites in the *N. benthamiana* genome.

Primers and gRNA sequences used

Locus	Forward 5'-3'	Reverse 5'-3'
Primers		
NbCYP82E2	TCCACTTCAATAACGACGGC	CGCCGTAAAGAAAAGCTGGA
LABCYP82E2 promoter	TTTAAATGGCCATATCAGAGATG	TTATGAATTTTTGGATAAGAATC
QLDCYP82E2 promoter	AAACCGCGGTTAAATGGCCATATCGGAG	AAACTCGAGTATGAATTTTTGGA TAAGAATC
NbGAPDH qPCR internal standard	CACTACCAACTGCCTTGCAC	ATGAAGCAGCTCTTCCACCT
pUQC-GFP-(218)-A Right Border 1	AACGCGCAATAATGGTTTCT	
pUQC-GFP-(218)-A Right Border 2	CCAAACGTAAAACGGCTTGT	
pUQC-GFP-(218)-A Right Border 3	CGCTCATGATCAGATTGTCG	
pFN117 Right Border 1	AATCCAGATCCCCCGAATTA	
pFN117 Right Border 2	CTGGCGTAATAGCGAAGAGG	
pFN117 Right Border 3	CGAATGCTAGAGCAGCTTGA	
Arbitrary degenarate primers for TAIL PCR (AD1)	NGTCGASWGANAWGAA	
Arbitrary degenarate primers for	TGWGNAGSANCASAGA	

TAIL PCR (AD2)		
Arbitrary degenarate primers for TAIL PCR (AD3)	AGWGNAGWANCAWAGG	
Arbitrary degenarate primers for TAIL PCR (AD6)	WGTGNAGWANCANAGA	
gRNA sequences used		
NbRDR1	TAAATAGTACAGTTTCTCCA	
	GACACTCAAAGTTTCTCTGG	
NbRDR2	CCACTCCCAACGTAGATAAG	
	GTGTCTCGAAATGTGCTGCA	
NbRDR6	CTTACTTAGAAGTCATCAGG	
	CTGCAACAGTATTACCAAAG	
NbPDS	TCACAAACCGATATTGCTGG	
	GAGCTTCAGGAAAATCAAAG	

Data availability

The *Nicotiana benthamiana* genome and transcriptome assemblies, along with their annotations, can be accessed at https://apollo.nbenth.com/. The raw data utilized for genome assembly has been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA881799. Specifically, the PacBio data for LAB and QLD can be found under the accessions SRR21820240 and SRR21820239, respectively. The HiC data for LAB and QLD are available under the accessions SRR21820238 and SRR21820237, respectively. Databases used: KEGG (https://www.genome.jp/kegg/compound/), Metfrag (https://ipb-halle.github.io/MetFrag/projects/metfragweb/), PubChem mass databases (ST3) (https://pubchem.ncbi.nlm.nih.gov/), miRbase (release 21; https://www.mirbase.org/) and Nicotiana attenuata Data Hub (http://nadh.ice.mpg.de/NaDH/others/data)

Code Availability

The code employed for obtaining chromosome-level genome sequences can be obtained from the following GitHub repository: https://github.com/anjiyuan/Citrus. The Circos plotter can be accessed via https://bioweb01.qut.edu.au/circos-bigwig/. Additionally, the synteny and dot plotter can be accessed through https://bioweb01.qut.edu.au/syntenyViewer/.

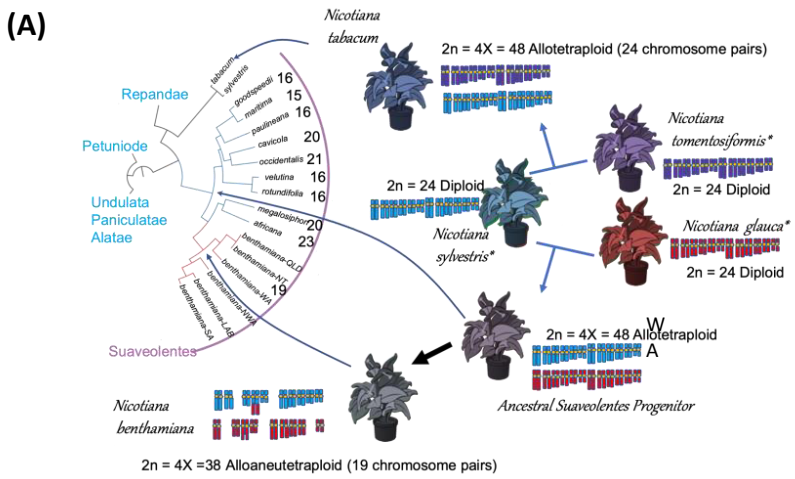
Methods References

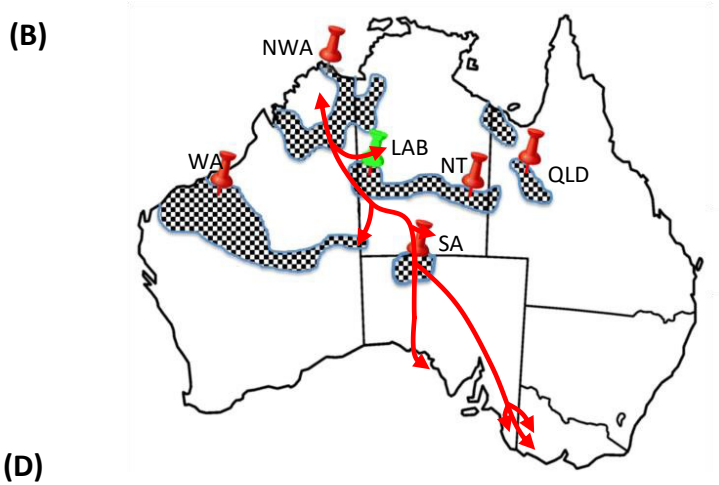
1. Bally, J. *et al.* The extremophile Nicotiana benthamiana has traded viral defence for early vigour. *Nat Plants* **1**, 15165 (2015).
2. Ruiz, M. T., Voinnet, O. & Baulcombe, D. C. Initiation and maintenance of virus-induced gene silencing. *Plant Cell* **10**, 937–946 (1998).
3. Philips, J. G. *et al.* The widely used Nicotiana benthamiana 16c line has an unusual T-DNA integration pattern including a transposon sequence. *PLoS One* **12**, e0171311 (2017).
4. Bombarely, A. *et al.* A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research. *Mol. Plant. Microbe. Interact.* **25**, 1523–1530 (2012).
5. Sulli, M. *et al.* An Eggplant Recombinant Inbred Population Allows the Discovery of Metabolic QTLs Controlling Fruit Nutritional Quality. *Front. Plant Sci.* **12**, 638195 (2021).
6. Ranawaka, B., Tanurdzic, M., Waterhouse, P. & Naim, F. An optimised chromatin immunoprecipitation (ChIP) method for starchy leaves of Nicotiana benthamiana to study histone modifications of an allotetraploid plant. *Mol. Biol. Rep.* **47**, 9499–9509 (2020).
7. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
8. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
9. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
10. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
11. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. Preprint at https://doi.org/10.1101/071282.
12. Ye, C., Ma, Z. S., Cannon, C. H., Pop, M. & Yu, D. W. SparseAssembler: de novo Assembly with the Sparse de Bruijn Graph. *arXiv* [cs.DS] (2011).
13. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci Rep* **6**: 31900.
14. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. **27**, 737–746 (2017).
15. Liu, J. *et al.* Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol.* **21**, 121 (2020).
16. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
17. Dudchenko, O. *et al.* De novo assembly of the genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
18. Dong, P. *et al.* 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Mol. Plant* **10**, 1497–1509 (2017).
19. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101 (2016).
20. Xu, G.-C. e *et al*. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**, (2019).
21. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
22. Howe, K. *et al.* Significantly improving the quality of genome assemblies through curation. *Gigascience* **10**, (2021).
23. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

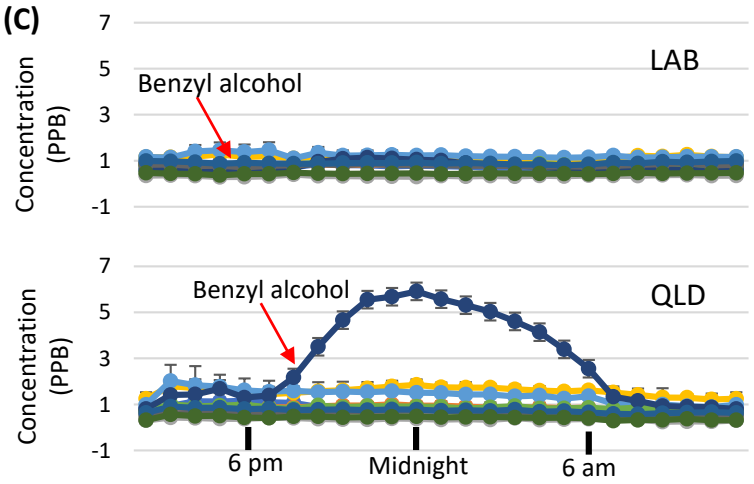
24. Shao, M. & Kingsford, C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* **35**, 1167–1169 (2017).
25. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33** (Web Server issue) W309-12.doi: 10.1093/nar/gkh379. (2005).
26. Dainat, J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. *Version v0* **4**, 10–5281 (2020).
27. 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).
28. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr Protoc* **1**, e323 (2021).
29. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol. 20*, 238 (2019).
30. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
31. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
32. An, J., Lai, J., Sajjanhar, A., Lehman, M. L. & Nelson, C. C. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. *BMC Bioinformatics* **15**, 275 (2014).
33. Huen, A., Bally, J. & Smith, P. Identification and characterisation of microRNAs and their target genes in phosphate-starved Nicotiana benthamiana by small RNA deep sequencing and 5'RACE analysis. *BMC Genomics* **19**, 940 (2018).
34. Baksa, I. *et al.* Identification of Nicotiana benthamiana microRNAs and their targets using high throughput sequencing and degradome analysis. *BMC Genomics* **16**, 1025 (2015).
35. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).
36. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
37. Lorenc, M. T. *et al.* Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology* **1**, 370–382 (2012).
38. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).
39. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).
40. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
41. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
42. Bandi V, Gutwin C, Siri JN, Neufeld E, Sharpe A, Parkin I. Visualization Tools for Genomic Conservation. *Methods Mol Biol.* **2443**, 285-308 (2022).
43. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
44. Siminszky, B., Gavilano, L., Bowen, S. W. & Dewey, R. E. Conversion of nicotine to nornicotine in Nicotiana tabacum is mediated by CYP82E4, a cytochrome P450 monooxygenase. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14919–14924 (2005).
45. Edgar, S. M. & Theriot, E. C. Phylogeny of Aulacoseira (Bacillariophyta) based on molecules and morphology1. *J. Phycol.* **40**, 772–788 (2004).
46. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).
47. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).

48. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
50. Singer, T. & Burke, E. High-throughput TAIL-PCR as a tool to identify DNA flanking insertions. *Methods Mol. Biol.* **236**, 241–272 (2003).

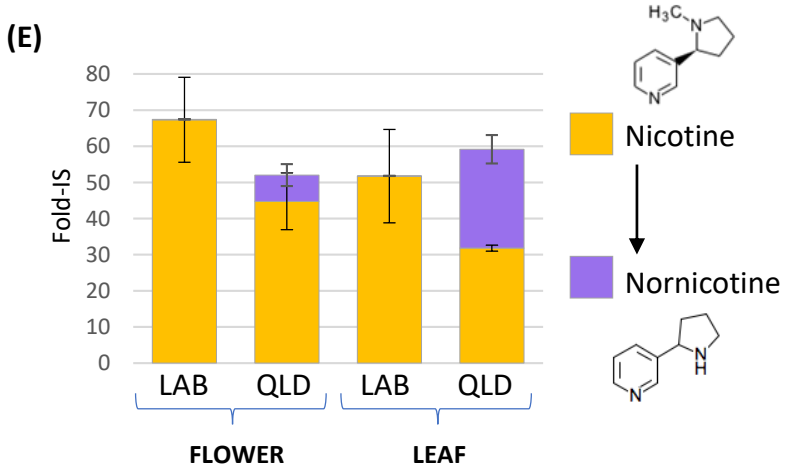
912
913
914
915
916
917
918
919
920
921
922

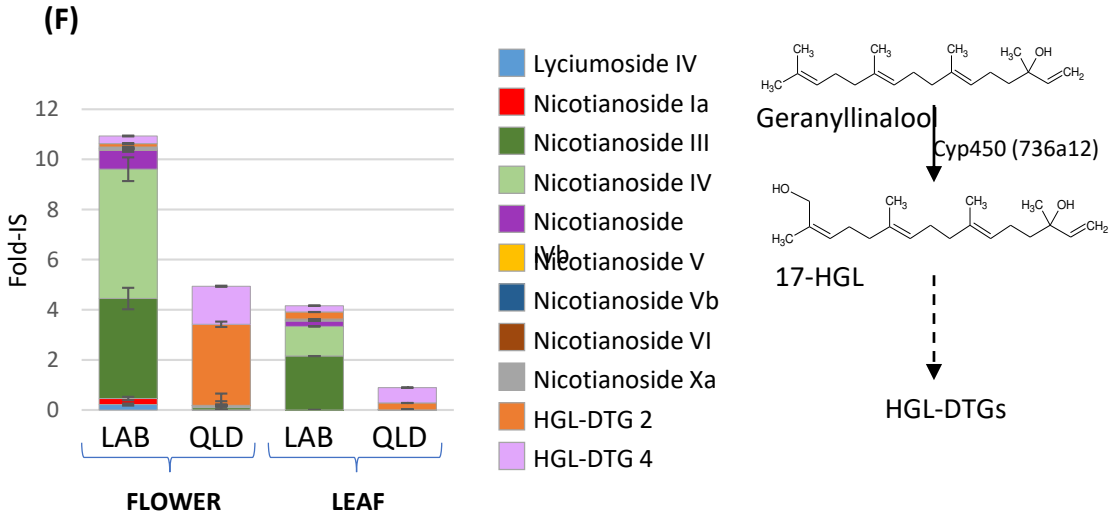

(A)
Nicotiana tabacum
2n = 4X = 48 Allotetraploid (24 chromosome pairs)
Repandae
Petuniode
Undulata
Paniculatae
Alatae
Suaveolentes
Nicotiana tomentosiformis*
2n = 24 Diploid
Nicotiana sylvestris*
Nicotiana glauca*
2n = 24 Diploid
2n = 4X = 48 Allotetraploid
W
A
Ancestral Suaveolentes Progenitor
Nicotiana benthamiana
2n = 4X =38 Alloaneutetraploid (19 chromosome pairs)

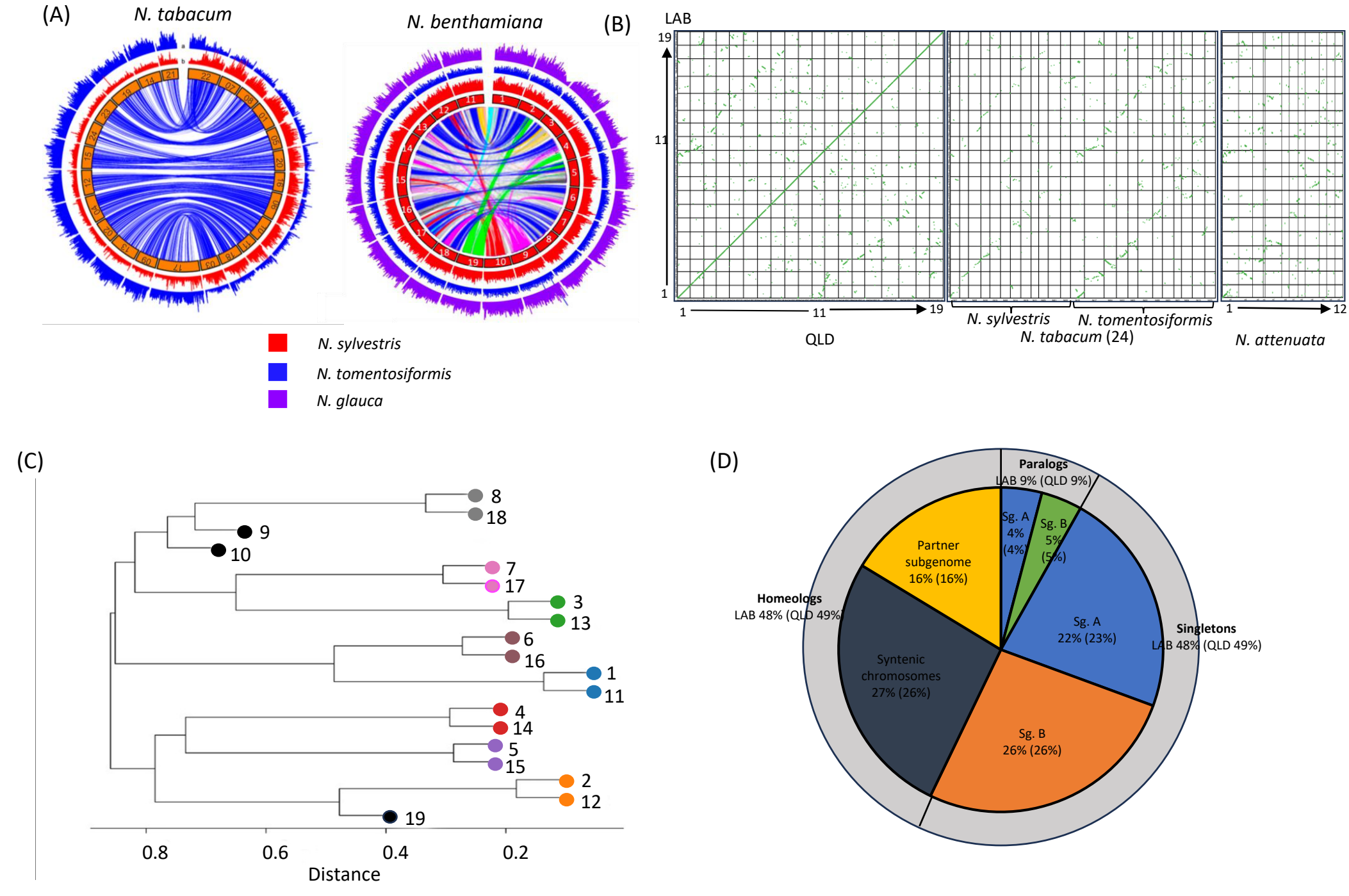

(B)
NWA
WA
LAB
NT
QLD
SA


(C)
Concentration (PPB)
LAB
Benzyl alcohol
QLD
6 pm
Midnight
6 am

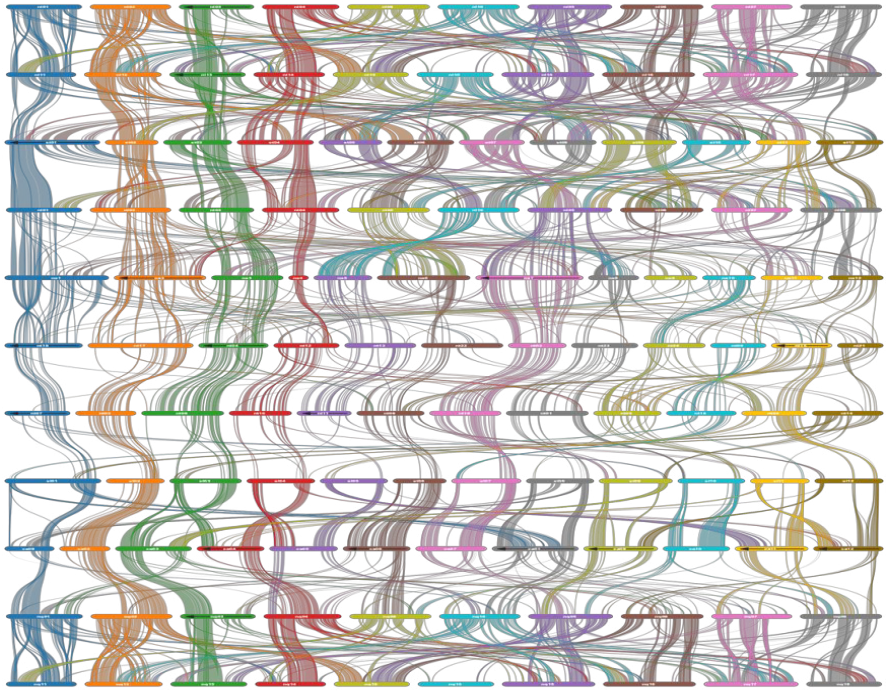

(D)


(E)
Fold-IS
LAB
QLD
LAB
QLD
FLOWER
LEAF
Nicotine
Nornicotine


(F)
Fold-IS
LAB
QLD
LAB
QLD
FLOWER
LEAF
Lyciumoside IV
Nicotianoside Ia
Nicotianoside III
Nicotianoside IV
Nicotianoside Vb
Nicotianoside V
Nicotianoside Vb
Nicotianoside VI
Nicotianoside Xa
HGL-DTG 2
HGL-DTG 4
Geranyllinalool
Cyp450 (736a12)
17-HGL
HGL-DTGs


(A)
N. tabacum
N. benthamiana
(B)
LAB
19
11
1
1
11
19
QLD
N. sylvestris
N. tomentosiformis
N. tabacum (24)
1
12
N. attenuata
N. sylvestris
N. tomentosiformis
N. glauca
(C)
8
18
9
10
7
17
3
13
6
16
1
11
4
14
5
15
2
12
19
0.8
0.6
0.4
0.2
Distance
(D)
Paralogs
LAB 9% (QLD 9%)
Sg. A
4%
(4%)
Sg. B
5%
(5%)
Partner
subgenome
16% (16%)
Homeologs
LAB 48% (QLD 49%)
Sg. A
22% (23%)
Singletons
LAB 48% (QLD 49%)
Syntenic
chromosomes
27% (26%)
Sg. B
26% (26%)

(A)



N.benthamiana LABa

N.benthamiana LABb

Potato

N.benthamiana LABa

N. attenuata

Tobacco$_{tom}$

Tobacco$_{syl}$

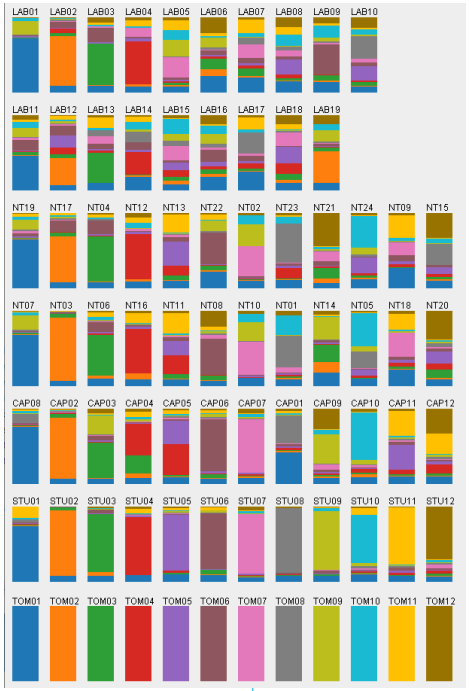
Tomato

Capsicum

N.benthamiana QLDa

N.benthamiana QLDb

(B)



N.b. LABa

N.b. LABb

Tobacco$_{tom}$

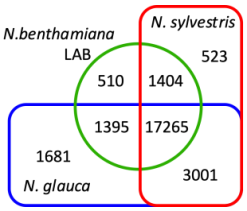
Tobacco$_{syl}$

Capsicum

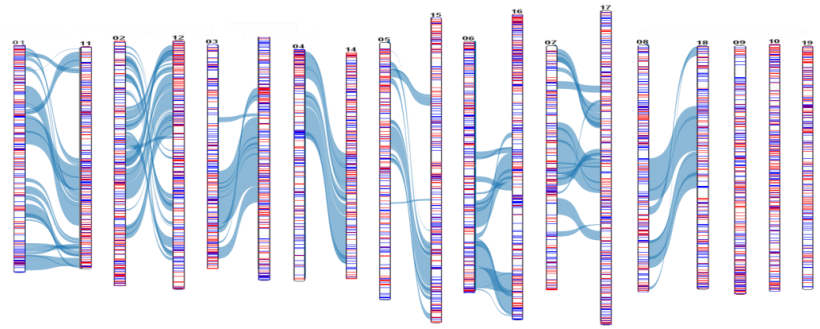
Potato

Tomato

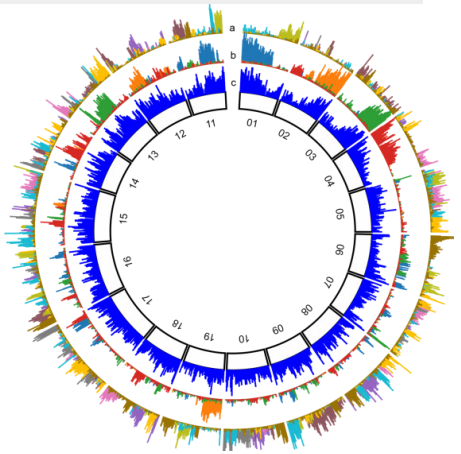
(C)



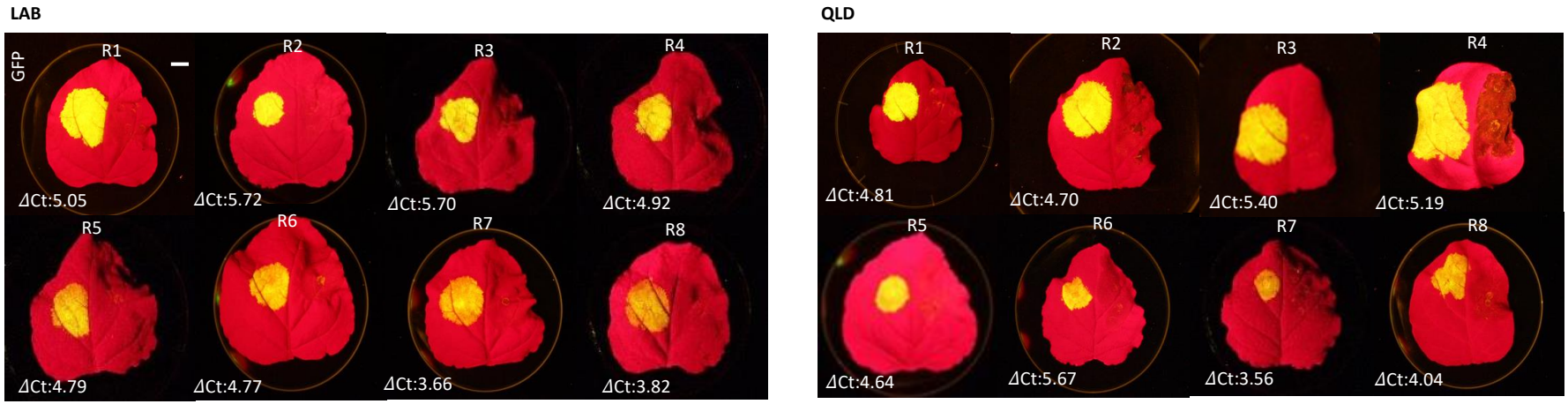
(D)



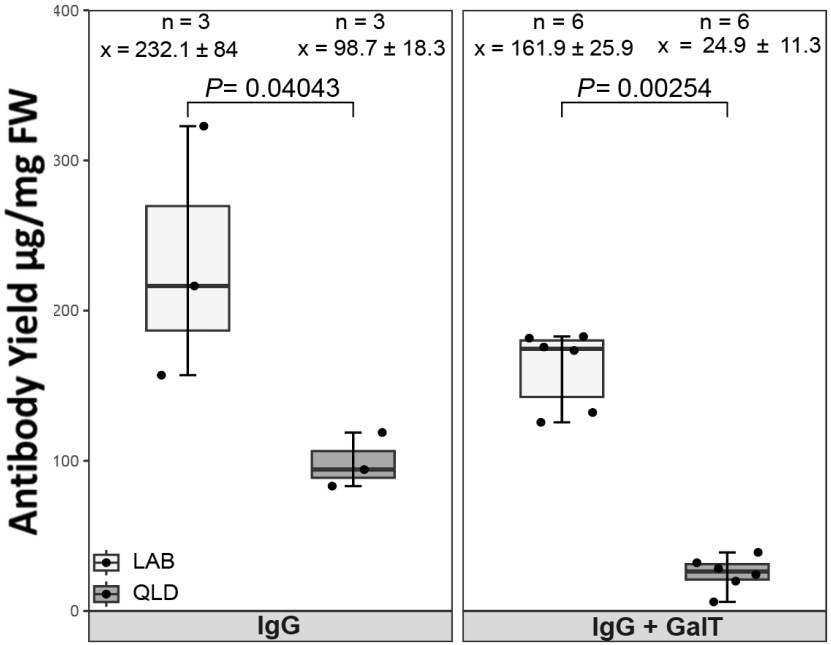
(E)



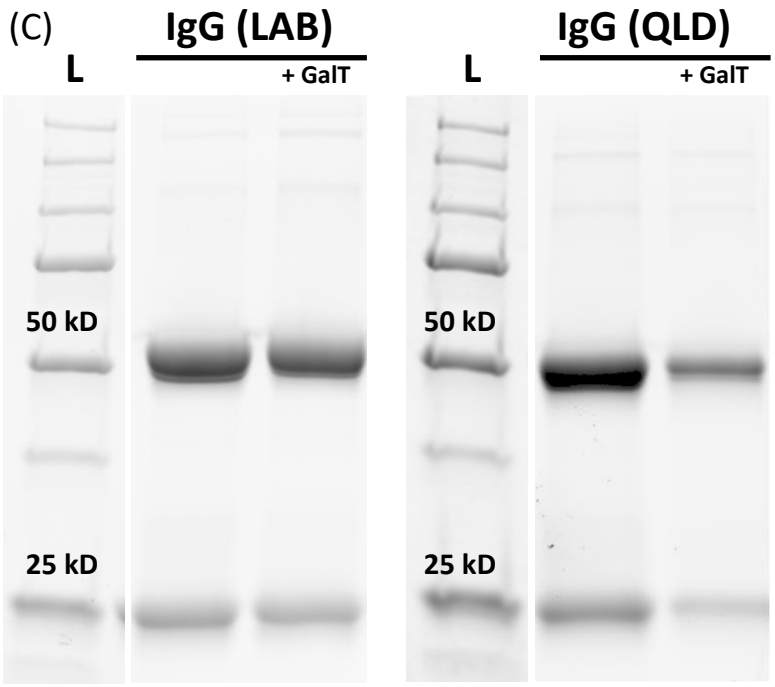
(A)

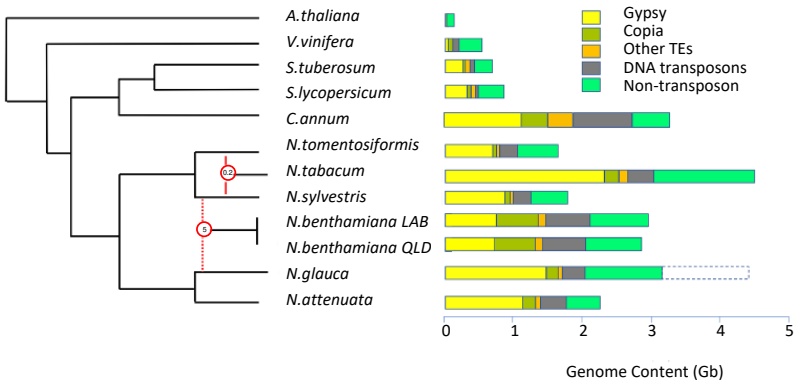

LAB
QLD
GFP
R1
R2
R3
R4
R5
R6
R7
R8
ΔCt:5.05
ΔCt:5.72
ΔCt:5.70
ΔCt:4.92
ΔCt:4.79
ΔCt:4.77
ΔCt:3.66
ΔCt:3.82
ΔCt:4.81
ΔCt:4.70
ΔCt:5.40
ΔCt:5.19
ΔCt:4.64
ΔCt:5.67
ΔCt:3.56
ΔCt:4.04

(B)

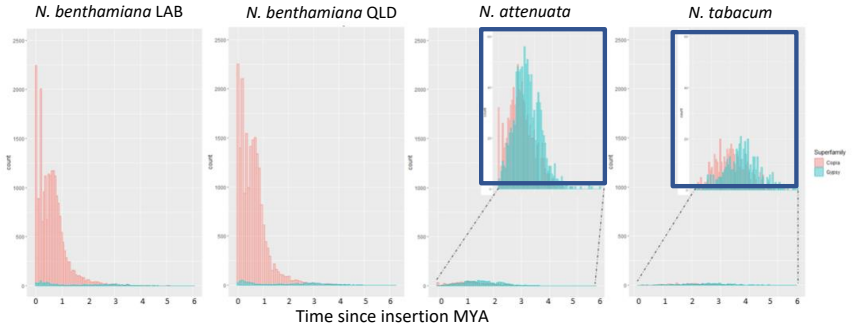

Antibody Yield µg/mg FW
400
300
200
100
0
n = 3
x = 232.1 ± 84
n = 3
x = 98.7 ± 18.3
P= 0.04043
n = 6
x = 161.9 ± 25.9
n = 6
x = 24.9 ± 11.3
P= 0.00254
LAB
QLD
IgG
IgG + GalT

(C)

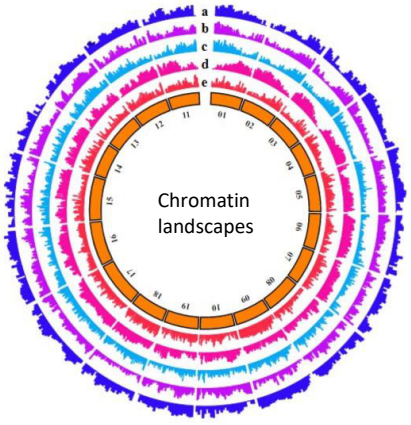

IgG (LAB)
IgG (QLD)
L
+ GalT
50 kD
25 kD


A.thaliana
V.vinifera
S.tuberosum
S.lycopersicum
C.annum
N.tomentosiformis
N.tabacum
N.sylvestris
N.benthamiana LAB
N.benthamiana QLD
N.glauca
N.attenuata
Gypsy
Copia
Other TEs
DNA transposons
Non-transposon
0
1
2
3
4
5
Genome Content (Gb)

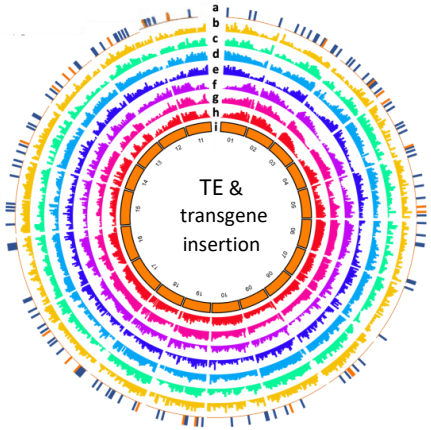
(A)


N. benthamiana LAB
N. benthamiana QLD
N. attenuata
N. tabacum
Time since insertion MYA

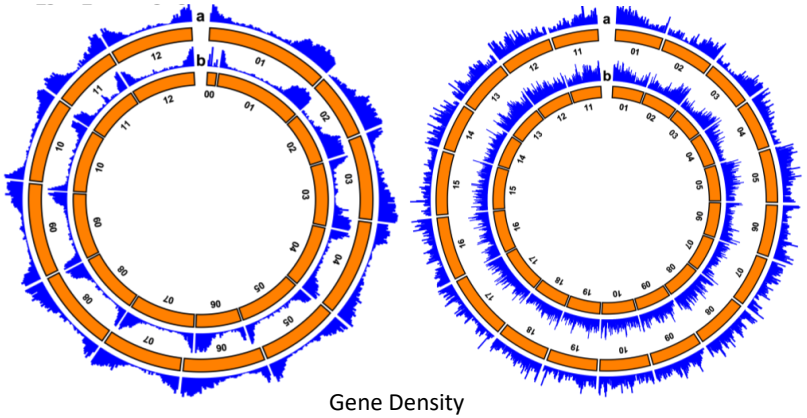
(B)


Chromatin
landscapes

(C)


TE &
transgene
insertion

(D)


Gene Density

(E)

(F)