NICOTIANA BENTHAMIANA: IN SILICO ANALYSIS OF ITS GENOME, GENE SPACE AND EVOLUTIONARY HISTORY

Written by:
Ph.D. candidate Matteo Schiavinato, M.Sc.

Institute of Computational Biology
group Himmelbauer

Department of Biotechnology
University of Natural Resources and Life Sciences (BOKU), Vienna

Supervisor
Univ.Prof. Mag. Dr.rer.nat. Heinz Himmelbauer

Co-supervisors
Univ.Prof. Dipl.-Ing. Dr. Lukas Mach
Ass.Prof. Dr. Juliane Dohm

Vienna, January 2020

BioToP Biomolecular Technology of Proteins

FWF

To everyone who thinks “I will never be good enough”.

Table of Contents

Abstract 6

Kurzfassung 7

Chapter 1: Introduction 9

- Background 10
- Sequencing data and plant genomes: a troubled story 12
- Genome assemblies: key concepts and issues in plants 14
- Read mapping in highly repetitive genomes 16
- Concepts of polyploidization 18
- *N. benthamiana* in research 21
- This project 22
- References 26

Chapter 2: Genome and transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔXT/FT 33

- Article 34
- Supplementary material 50

Chapter 3: Parental origin of the allotetraploid tobacco *Nicotiana benthamiana* 75

- Article 76

Chapter 4: Conclusions 107

Chapter 5: Appendix 110

- List of publications 111
- Curriculum Vitae 112
- Acknowledgements 113

Abstract

Nicotiana benthamiana is an allotetraploid species of tobacco that is native to Australia. It belongs to the *Solanaceae* family, together with many other plants with relevance to agriculture as well as to the life sciences. The origin of its hybrid genome is still debated, with literature supporting different maternal progenitors and hybridization dates. In the last decades, *N. benthamiana* was used extensively to produce recombinant glycoproteins *in planta* and to study host-pathogen interactions, and the interest around this plant has grown substantially. Despite of this, prior to this work, the molecular resources available for this plant were limited, even though two genome assemblies and one transcriptome assembly had already been reported. Nevertheless, a multitude of sequencing data points had accumulated in public repositories.

In this thesis I made use of public and newly generated high-throughput cDNA sequencing data to re-analyse the genome of *N. benthamiana*. I calculated a gene set for *N. benthamiana* using an approach which combined *in silico* prediction with cDNA-based transcript verification. Thereafter, the gene set was used to perform several downstream analyses. I performed a differential gene expression study in one research line of *N. benthamiana* and characterized the insertion site of a transgene in its genome. I analysed the inter-accession variation in research lines used in different laboratories around the world, showing that they likely all originate from one original source. I generated a large collection of phylogenetic trees encompassing all *N. benthamiana* genes and their homologs in six other *Nicotiana* species. I used it to study the evolutionary history of the *N. benthamiana* genome, addressing its parental progenitors and its hybridization date. I show that the hybridization is likely to have taken place around five million years ago between ancestors of *Nicotiana* section *Noctiflorae* (as maternal parent) and *Sylvestres* (paternal parent).

Kurzfassung

Nicotiana benthamiana ist eine allotetraploide Tabakart aus Australien. Die Art gehört zu den Solanaceen, zusammen mit anderen Pflanzen mit Relevanz entweder für die Landwirtschaft oder für die Biowissenschaften. In der Literatur gibt es unterschiedliche Angaben zur möglichen mütterlichen Elternspezies von *N. benthamiana* und zum Hybridisierungszeitpunkt. In den letzten Jahrzehnten wurde *N. benthamiana* zur Herstellung rekombinanter Glykoproteine *in planta* und zur Untersuchung von Wirt-Pathogen-Interaktionen verwendet. Dadurch ist das Interesse an dieser Pflanze erheblich gewachsen. Trotzdem waren die für diese Pflanze verfügbaren molekularen Ressourcen bisher begrenzt, obwohl bereits zwei Genomsequenzen und ein assembliertes Transkriptom veröffentlicht worden waren.

In dieser Arbeit nutzte ich öffentlich verfügbare und neu generierte cDNA-Sequenzen aus der Hochdurchsatzsequenzierung, um das Genom von *N. benthamiana* zu untersuchen. Ich berechnete einen Gendatensatz für *N. benthamiana* mit einem Verfahren, das *in silico*-Vorhersage mit cDNA-basierter Transkript-Verifizierung verknüpfte. Mit dem Gendatensatz wurden mehrere Analysen durchgeführt: Im Vergleich zu einer Kontrolle ermittelte ich differentiell exprimierte Gene in einer Akzession von *N. benthamiana* und charakterisierte eine Transgen-Insertionsstelle in ihrem Genom. Ich analysierte die Variation zwischen Akzessionen aus verschiedenen Laboratorien und zeigte, dass sie wahrscheinlich aus einer einzigen Quelle stammen. Ich errechnete eine große Anzahl an phylogenetischen Bäumen, die einen Großteil der Gene von *N. benthamiana* repräsentierten und ihre Homologen in sechs anderen *Nicotiana*-Arten umfassen. Anhand dieser Bäume wurde die Evolutionsgeschichte des *N. benthamiana*-Genoms untersucht. Ich konnte zeigen, dass die Hybridisierung wahrscheinlich vor etwa fünf Millionen Jahren zwischen Vorfahren der *Nicotiana*-Sektion *Noctiflorae* (mütterlicher Elternteil) und *Sylvestres* (Elternteil väterlicherseits) stattfand.

Chapter 1:

Introduction

Background

Nicotiana benthamiana is a species of tobacco indigenous to Australia. The genus *Nicotiana* is named after Jean Nicot, a French scholar of the 16th century who was the first to import tobacco plants from Portugal to France; the species name *benthamiana*, instead, is a homage to the English botanist George Bentham (Goodspeed, 1954).

The *Nicotiana* genus belongs to the *Solanaceae* family, which features some of the most relevant species in agriculture (Figure **1**), such as tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), the smoking tobacco (*Nicotiana tabacum*), bell pepper (*Capsicum annuum*) and eggplant (*Solanum melongena*) (Bally et al., 2018). The *Solanaceae* family also includes several ornamental plants, among which several tobacco species, and plants used in traditional medicine such as *Atropa belladonna*. Overall, most species belonging to the *Solanaceae* have been domesticated to exploit their products, becoming economically relevant in agriculture, research and medicine.

Figure 1: from **Wang L, Li J, Zhao J and He C** (2015) Evolutionary developmental genetics of fruit morphological variation within the Solanaceae. Front. Plant Sci. 6:248. doi: 10.3389/fpls.2015.00248

The diverse variations of fruit morphology in the Solanaceae family. (1–3), *Solanum melongena*; (4), *Solanum pimpinellifolium*; (5–8), *Solanum lycopersicum*; (9–14), Variants of *Capsicum annum*; (15), *Physalis alkekengi*; (16), *Physalis floridana*; (17–19), *Physalis philadelphica*. The Chinese lantern in *Physalis* spp. was opened to show the berry inside. Bar = 1 cm.

The most prominent member of the *Nicotiana* genus is *Nicotiana tabacum*, the smoking tobacco, for which multiple genomic and proteomic resources are available (Edwards et al., 2017; Sierro et al., 2014). *N. tabacum* has also been extensively used as expression platform, i.e. to produce recombinant proteins in large scale (Conley et al., 2011). The same can be said for many other *Nicotiana* species, including *N. benthamiana* (Wydro, Kozubek, & Lehmann, 2006). However, contrary to *N. tabacum*, molecular resources available for *N. benthamiana* are scarce. Throughout the last decade it has been used extensively for the production of recombinant proteins (Jansing, Sack, Augustine, Fischer, & Bortesi, 2018; Joensuu et al., 2010; J. Li et al., 2016; Montero-Morales et al., 2017; Strasser et al., 2008; van Herpen et al., 2010), and it has also been exploited as model organism to study host-pathogen interaction (Bally et al., 2018; Goodin, Zaitlin, Naidu, & Lommel, 2008; Matsumura et al., 2003). Despite its growing usage, when this work began, the genome and transcriptome of this species were not well characterized. Two draft genome sequences were available (Bombarely et al., 2012; Naim et al., 2012), which had been generated to aid gene editing and genome targeting. Both sequences came with an associated predicted set of genes from *in silico* predictions; a transcriptome assembly was also generated few years later (Nakasugi, Crowhurst, Bally, & Waterhouse, 2014). However, these gene sets did not live up to expectations due to the lack of integration with experimental data from cDNA sequencing, which can result in many wrongly predicted models. In the following years, due to a rapid increase in interest for *N. benthamiana*, several new projects began to funnel data into public repositories (Bally et al., 2015; Alexandra Castilho et al., 2015; J. Li et al., 2016; Long, Ren, Xiang, Wan, & Dong, 2016), many of which providing useful sequencing data. At this point, a well-supported gene prediction became possible and timely, together with an in-depth analysis of the transcriptome, the proteome and the phylogeny of the species.

Addressing the genome properties of *N. benthamiana* has been demanding due to its highly repetitive polyploid genome (60%, see Schiavinato et al., 2019). In fact, the repetitive content of a genome poses a great challenge for any *in silico* method based on sequencing reads. In the following introductory paragraphs, key concepts of plant genomes are described, as well as methodological aspects that are relevant to this work, to better understand the context of this thesis.

Sequencing data and plant genomes: a troubled story

Plant genomes have a very low gene density compared to other eukaryotes. One of their key features is, in fact, their enrichment in repetitive DNA and transposable elements, which contribute sensibly to the non-coding genome portion (Bennetzen, 2000; Hanson et al., 1998; K.Y. Lim, Matyasek, Kovarik, Fulnecek, & Leitch, 2005; B. C. Meyers, 2001; Petit et al., 2010). Repetitive DNA consists of stretches of simple sequences that are repeated multiple times within the genome. This type of DNA is highly enriched in homopolymers which can induce unequal crossing over, resulting in a duplication of the region. Transposable elements (TEs), instead, are more organized stretches of DNA who have the ability to move within a genome (Bennetzen, 2000; Flavell, Pearce, & Kumar, 1994; Kejnovsky, Hawkins, & Feschotte, 2012). The mechanism with which TEs move within a genome is similar to that of retroviruses (Grandbastien et al., 1989): In fact, it is thought that TEs are leftover DNA from retroviral infections, but the viral DNA does not encode the viral coat proteins anymore.

TEs are the most important class of repetitive elements in plants. They were first characterized by Barbara McClintock in maize, where they were referred to as “controlling elements” due to their ability to control certain phenotypes and to trigger a “genomic shock” (McClintock, 1984). The main feature of TEs is their ability to move elsewhere in the genome, triggering an unpredictable cascade of events. Two classes of TEs exist based on their transposing strategy (Figure **2**). Class I TEs are transcribed into RNA and then retro-transcribed to a dsDNA fragment that can insert in another part of the genome. This increases their copy number and expands the size of the genome. The expansion of a genome through TEs can happen rapidly and in large amounts, a phenomenon called “TE burst” (Laudencia-Chingcuanco & Fowler, 2012). Class II TEs, instead, use a different strategy: they depend on a transposase gene, either encoded by themselves or by a nearby transposon. The transposase can cleave their flanking sequences, excising them from their current genome position and allowing them to insert somewhere else (Flavell et al., 1994). This does not increase the size of a genome, but can lead to gene disruptions and, ultimately, phenotypic variations.

Repetitive DNA and TEs can play a role in development, in gene regulation and may be responsible for certain phenotypic variations. Although present in all eukaryotic species (Biscotti, Olmo, & Heslop-Harrison, 2015), their contribution to plant genome size outweighs that in any other eukaryote. Given their abundance they are also used to determine phylogenetic relationships (K.Y. Lim et al., 2005). In fact, the same element could vary in copy number and may accumulate point mutations between different species.

When sequencing a plant genome, all these types of non-coding DNA are challenging. Homopolymers in repetitive DNA are a well-known issue for sequence determination, leading to multiple artefacts and false basecalls (Dohm, Lottaz, Borodina, & Himmelbauer, 2008; Lu,

Giordano, & Ning, 2016; Minoche, Dohm, & Himmelbauer, 2011; Quail et al., 2012; Ross et al., 2013). They are also difficult to handle in read mapping, the reasons of which are explained in one of the following paragraphs. Moreover, TEs must be carefully taken into account when performing gene predictions, as they by and large resemble genes and might trick the gene prediction algorithm into believing that a gene locus is present (Minoche et al., 2015). All in all, it is important to remember that the repetitive content of a plant genome is likely to introduce an issue in any stage of its analysis.

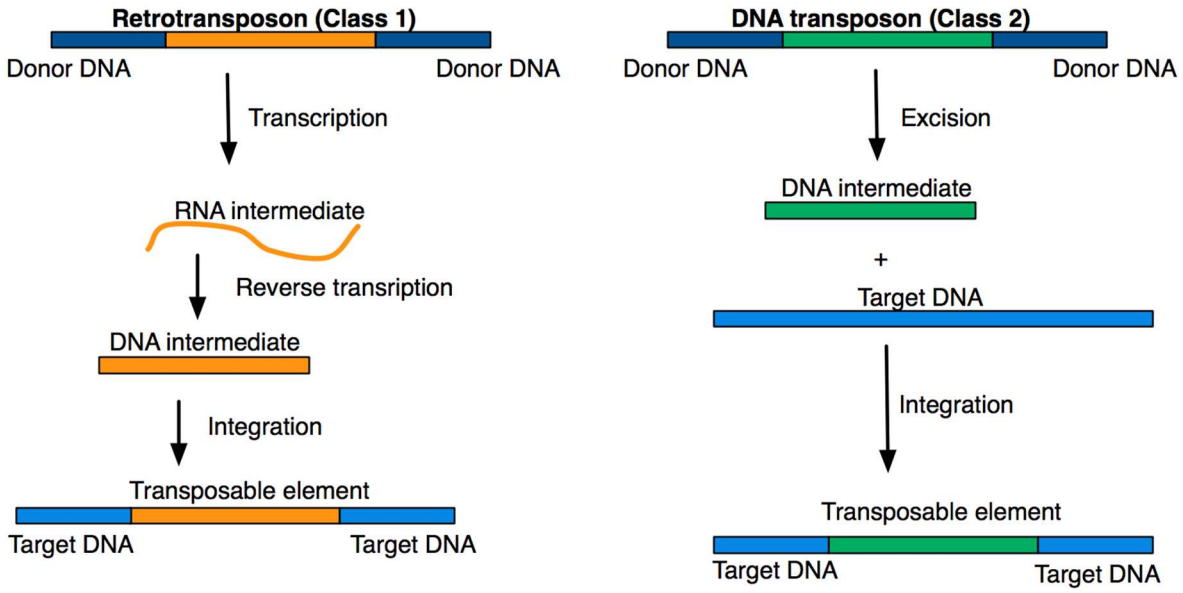


Figure 2: from Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Transposable_elements_(2).png

Classification of transposable elements. Transposable elements self-replicate through two main mechanisms: via an RNA intermediate ("copy-and-paste"; class 1) or straight excision-insertion ("cut-and-paste"; class 2).

Genome assemblies: key concepts and issues in plants

An assembly approximates a genome sequence. It is a collection of sequences, called contigs, each representing a portion of the sequenced genome. Contigs are obtained by merging overlapping sequencing reads, and are organized into scaffolds (Imelfort & Edwards, 2009). Scaffolds represent 'ordered' contigs, the order of which is determined using information from multiple sources: physical maps (Blake C. Meyers, Scalabrin, & Morgante, 2004), optical maps (Aston, Mishra, & Schwartz, 1999) or mate pairs (Chaisson, Brinza, & Pevzner, 2008). The number of scaffolds might range from tens to thousands, depending on the amount of data used, on the assembly quality and on the genome complexity. A schematic is shown in Figure **3**.

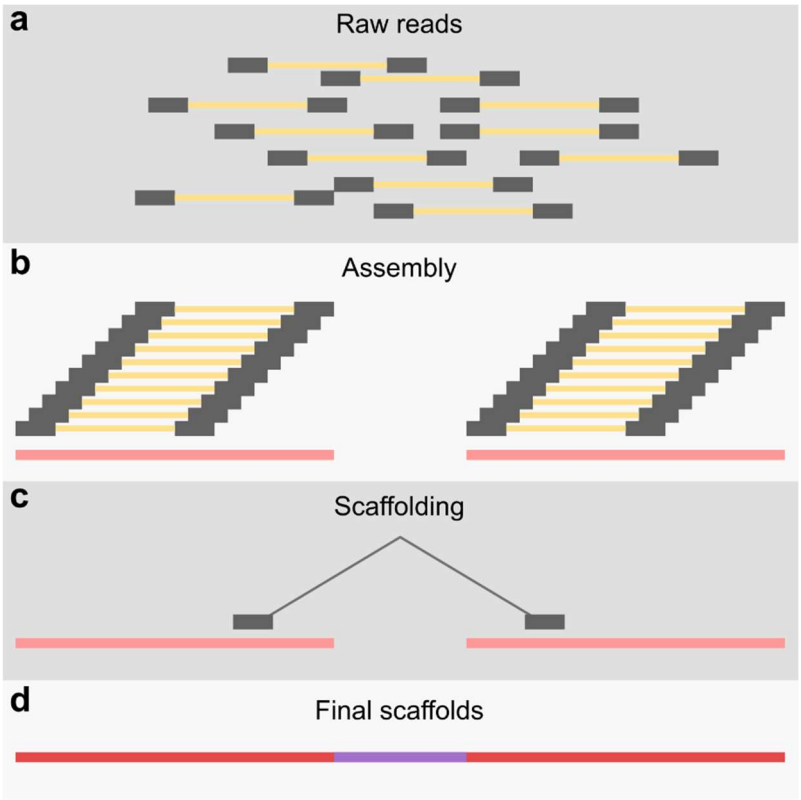


Figure 3: *de novo* genome assembly workflow. a) Raw reads are generated on a sequencing machine. Shown are Illumina paired-end short reads. Black bars represent sequenced portions, yellow lines show the connecting un-sequenced DNA. **b)** Overlapping reads are grouped together and merged into contigs (light red bars). **c)** Paired-end reads generated from larger fragments (mate pairs) are used to detect physical connections between existing contigs. **d)** Connected contigs are joined with a stretch of Ns (violet bar). The connected contigs form a scaffold (dark red bars).

The genomes of the agriculturally most relevant plants have been generated with a very high amount of data from multiple sources (Bertioli et al., 2019; Edwards et al., 2017; Jarvis et al., 2017; F. Li et al., 2015; Velasco et al., 2010). Thus, the number of scaffolds contained in those assemblies parallels the number of chromosomes. However, having one scaffold per

chromosome is mostly an ideal case and does not reflect the majority of published plant genome assemblies.

When sequencing a plant genome, many produced sequencing reads are identical or extremely similar due to the genome's repetitive content. When these reads are used to make an assembly, they will lead to several known issues. On the one hand, copy number information is lost due to the merging of identical (or very similar) reads. On the other hand, they might trick the genome assembler into joining sequences that are, in fact, distant in the genome. Genome assembly tools tackle this issue by breaking the nascent assembled sequences at these locations (Luo et al., 2012). However, this control mechanism doesn't always work, and when it does, the resulting assembly is highly fragmented. Hence, in many plant genomes a chromosome-level assembly is only an ideal scenario that might be economically out of reach for many research groups. Next-generation sequencing reads are too short to entirely overcome this obstacle. The emerging long-read technologies such as those from Pacific Biosciences (Eid et al., 2009) and Oxford Nanopore (Jain, Olsen, Paten, & Akeson, 2016) are providing partial improvements, but their basecalling reliability is inferior to the one of short read technologies and is still sensitive to homopolymers such as tandem repeats. The length of their reads is in the range of the thousand base pairs, often spanning an entire stretch of repetitive genomic DNA. In many cases this is enough to obtain a higher quality assembly, but the problem is far from being solved. All in all, it is important to stress that genome assemblies do contain errors and artefacts, often overlooked by researchers who only use a genome sequence for specific studies.

Read mapping in highly repetitive genomes

A second issue when dealing with plant sequencing data comes into play in read mapping experiments, due to reads originating from repetitive regions. Before explaining the issue, however, here are some key concepts about read mapping.

The output of a short-read sequencing reaction (such as Illumina sequencing) contains usually millions of reads, each around 100 nt long. Long-read sequencing machines, instead, produce fewer reads but they are kilobases in length. In both cases, a read represents a stretch of DNA (or cDNA) from the sequenced genotype, which can then be compared to any other genotype to extract a various amount of information.

Read mapping is an operation in which a sequencing read is compared to a reference sequence to find the position where it is derived. A sequencing read is a sequence of nucleotides, and so is the reference sequence. The mapping location of a read is the position in the reference sequence that has the highest sequence identity to the read. This is where the issue emerges: reads from repetitive elements will map to multiple locations with comparable (if not identical) quality, so it is hard to determine from which of these many regions a read comes from. A workaround used in many studies, including ours (Dohm et al., 2013; Minoche et al., 2015; Schiavinato et al., 2019), is that of "masking" the transposable elements in assembled genomes (Smit & Hubley, 2008; Smit, Hubley, & Green, 2013). TE positions in the reference genome are annotated and reads mapped within them are filtered out of the analysis. This solution might be cleaner for variant calling and other sequencing-based analyses performed in coding and intergenic DNA. However, one must keep in mind that more than 50% of a plant genome is usually repetitive (Bennetzen, 2000), and by masking it, it is completely disregarded.

To find the mapping location with the highest identity to the read, most popular mapping algorithms use a mapping strategy called seed-and-extend (Figure **4**), which trades off part of the mapping specificity in favour of a substantial gain in computational speed (Kim, Langmead, & Salzberg, 2015; Langmead & Salzberg, 2012; H. Li & Durbin, 2010). This strategy is based on an index, which is generated on the target genome sequence prior to the alignment. The details of the indexing are beyond the scope of this thesis and won't be included. What is relevant is that an index contains a series of short sequences and their location in the genome, which serves as a look-up table for mapping tools to find candidate mapping locations for the reads. When a read is mapped, the algorithms extract from the read a random sequence of length *k* (usually around 20) called "seed" and looks up the genome index to find where the seed is found in the genome. Each seed is then "extended", matching the rest of the read to the genome. During the extension, a series of mismatches and gaps are often found. These could either be originating from sequencing errors or from biological differences between the

genotype of the reads and that of the reference sequence. Mapping tools allow the users to specify a maximum number of mismatches and gaps for each read; all seeds that exceed these thresholds during extension are no further considered. The length and the identity of the extension define the mapping quality and the mapping score, which are criteria used to rank alignment records. From the ranked records, the best alignment for each read is usually chosen (referred to as primary alignment). All the other alignments are referred to as "secondary", and while they still contain valuable information for certain analyses, most of the times they are disregarded.

In cases where most seeds find a unique location in the index, the alignment is trivial. When many locations are found, instead, each must be extended, which can add up to a non-negligible amount of time. It could happen that several locations produce equally good alignments, contributing to the background noise. Hence, the presence of several repetitive elements (usually above 50% of the whole genome) leads to many genomic reads assigned to multiple positions. The choice of a primary alignment can be made ranking the mapping records by mapping quality, or by alignment length; however, when differences are minimal, the choice is somewhat arbitrary and different algorithms might take different decisions. This sensibly increases artefacts, computational times and resources (Treangen & Salzberg, 2012). Hence, genomes with a high repetitive content pose a great challenge in this respect.

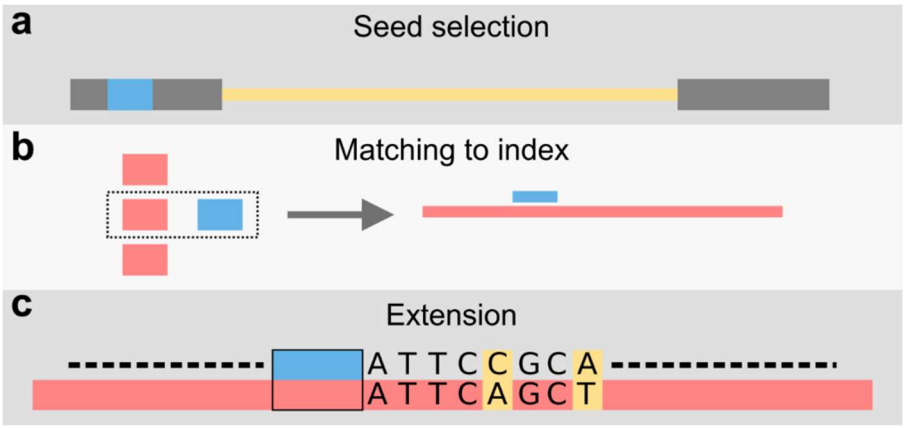


Figure 4: Seed-and-extend alignment strategy. a) A seed of length *k* is selected from a paired-end read (blue bar). **b)** The seed is compared to the genome index, finding all its locations onto the reference sequence (red bars, left of the arrow). Each found location is then sent to alignment (red bar, right of the arrow). **c)** The alignment is extended on both sides of the seed, comparing each read position (above line) with the paired reference position (below line, onto the red bar). Mismatches can be found (yellow columns).

Concepts of polyploidization

The history of flowering plants is heavily characterized by polyploidization events (Soltis & Soltis, 2009). Even *Arabidopsis thaliana,* which has one of the smallest plant genomes (0.157 Gbp) contains traces of ancient polyploidization (Vision, Brown, & Tanksley, 2000). Two main types of polyploidization exist: whole genome duplication (i.e. autopolyploidization) and hybridization (i.e. allopolyploidization); the latter has been made responsible for many speciation events, including the one of *N. benthamiana*. A hybridization event between two different species produces an amphidiploid organism containing both parental genomes within the same nucleus. In hybrid genome terms, these two genomes are referred to as the hybrid "subgenomes" (Soltis, Marchant, Van de Peer, & Soltis, 2015). The subgenomes are often quite closely related and many (if not all) chromosomes of one subgenome can be found in the other one too. This peculiar situation defines a specific case of paralogy: genes found in both subgenomes are paralogs (contained within the same genome) but originally orthologs (deriving from two independent species). The chromosomes containing these genes are referred to as "homeologs". As will be shown further in this work, the distinction between homeologs of a hybrid genome is not trivial and is at the ground of many issues when dealing with plant genomic data.

The concept of 'hybrid' is often debated, and it is hard to draw a line between hybrid and autopolyploid. The reason is that the difference between species and populations is sometimes quite narrow (Grant, 1981; Levin, 2000). One of the most successful definitions of species starts from an evolutionary point of view: two populations are considered as two separate species if they evolve separately and don't merge (Wiley, 1978). Occasional hybridization events are tolerated, if they don't bring the entire populations together. A hybrid is defined as a species generated from two parents whose genomes differ in at least one heritable character (Harrison & others, 1990).

Hybridization is considered as a positive speciation mechanism, a "stimulus driving evolution" (Anderson & Stebbins, 1954). Speciation after polyploid formation is quite common, since the doubling of the genomic material tends to prevent backcrossing of the hybrid with its original parents (Grant, 1981). From an evolutionary standpoint, polyploidization creates a sudden excess of genetic material, which could either be discarded or repurposed through neo- and sub-functionalisation mechanisms (Cheng et al., 2018). Different species might adopt either one of the strategies, often depending on the environment and the selective pressure that their niche imposes. In any case, the repurposing and the loss of the DNA in excess brings the polyploid genome back to a semi-diploid state (Figure **5**). This dynamic is known as "Long Term Genome Diploidization" (LTGD) and has been observed in many plant genera, including *Nicotiana* (Leitch et al., 2008). During LTGD, homoeologous recombination between the two

subgenomes may also take place (Canady, Ji, & Chetelat, 2006; Glover, Redestig, & Dessimoz, 2016; Udall, Quijada, & Osborn, 2005).

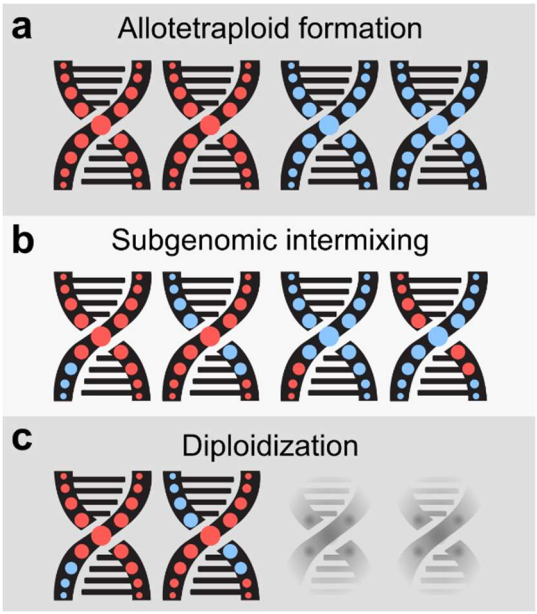


Figure 5: schematic representing subgenomic intermixing and Long-Term Genome Diploidization (LTGD). a) Two diploid genomes (red and blue) fuse together, forming an amphidiploid nucleus. This is the earliest stage of a hybrid genome formation. The two genomes are now subgenomes of an allotetraploid genome. **b)** The two subgenomes exchange genetic material in homoeologous recombination (unmatched coloured dots within chromosomes). **c)** Genetic material is lost (genome downsizing) in a tendency to bring the genome back to a diploid state (shown with the smaller size of the two right chromosomes).

Polyploid species that are subjected to LTGD are broadly classified into paleo- (> 30 million years), meso- (5-30 million years) and neo-polyploid (0-5 million years). Neo-polyploids usually still have the exact sum of the chromosomes of their parents (Bertioli et al., 2019). Meso-polyploids usually show a more visible downsizing, with the loss of certain chromosomes (Clarkson et al., 2005). Paleo-polyploids are often now diploid plants that carry traces of ancient polyploidization events (Julca, Marcet-Houben, Vargas, & Gabaldón, 2018; Renny-Byfield, Gong, Gallagher, & Wendel, 2015; Renny-Byfield et al., 2013). The hybrid genome of *N. benthamiana* has a haploid set of n=19 chromosomes (Bally et al., 2018). This number does not correspond to the exact sum of the haploid sets of the presumable parents (both having n=12). It is suspected that the genus *Nicotiana* is particularly subjected to genome downsizing and LTGD (Leitch et al., 2008; K. Yoong Lim et al., 2007), and that the number of chromosomes in *N. benthamiana* reflects this.

Many *Solanaceae* plants are polyploids (Bombarely et al., 2016). The genus *Nicotiana* is subdivided into several sections (Figure **6**, taken from Schiavinato et al., in press), grouping together individuals with shared genomic and morphological traits (Goodspeed, 1954; Knapp, Chase, & Clarkson, 2004).

A broad subdivision is the one between diploid and hybrid sections. Among the polyploid sections there is *Suaveolentes*, which contains *N. benthamiana* (Goodin et al., 2008).

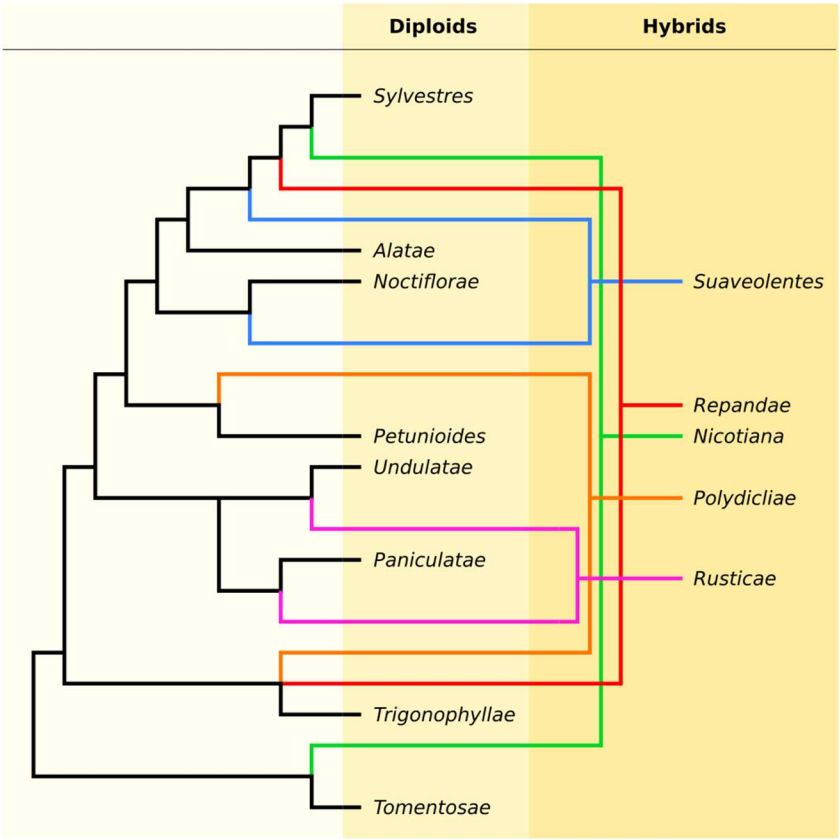


Figure 6: Representation of the *Nicotiana* sections and hybrids. Tree branches are intended to show relationships but are not scaled to actual phylogenetic distance. The tree topology is based on a previous study **(Knapp et al., 2004)**. Black branches indicate the evolution of diploid taxa, coloured branches refer to the evolution of hybrid taxa.

Suaveolentes plants originated from a hybridization event between two diploid *Nicotiana* species, which (prior to this work) was believed to have taken place between 6 and 10 million years ago (Bally et al., 2018; Clarkson, Dodsworth, & Chase, 2017). The two parental progenitors hybridized in South America and later diffused to Australia, Africa and the Pacific islands (Clarkson et al., 2017). Its paternal progenitor has been since long attributed to an ancestor of the extant species *N. sylvestris*, belonging to the homonymous section *Sylvestres* (Goodin et al., 2008). Its maternal progenitor, instead, has been debated for long time. Multiple candidates have been proposed in the literature, the most prominent ones being ancestors of sections *Noctiflorae* or *Petunioides* (Bally et al., 2018; Clarkson et al., 2017; Goodin et al., 2008; Kelly, Leitch, Clarkson, Knapp, & Chase, 2012; Leitch et al., 2008). In general, the evolution of hybrid polyploid genomes is characterized by strong genome mobility, downsizing and rearrangement, as well as substantial gene loss. This reduces the power of the assumptions that can be made on genome synteny and gene homology. Some of the challenges faced in hybrid genome analyses are described in the two publications included in this thesis.

N. benthamiana in research

It is known that *N. benthamiana*'s genome exchanged the loss of viral defence mechanisms for early vigour (Bally et al., 2015), a trade-off that conferred it a great genome plasticity. In fact, contrary to most of the other *Solanaceae* species, *N. benthamiana* is not a relevant food crop but rather a popular host for recombinant protein production, and a model organism for host-pathogen interaction (Fischer & Emans, 2000; Schillberg, Emans, & Fischer, 2002; van Herpen et al., 2010). Its popularity arose in the last two decades, due to its high susceptibility to agroinfiltration, which highlighted its potential as expression platform for recombinant proteins (Kościańska, Kalantidis, Wypijewski, Sadowski, & Tabler, 2005; Liu et al., 2010; Montero-Morales et al., 2017). This specific trait has been exploited to produce antibodies in large scale (A Castilho et al., 2011), also leveraging modern gene editing techniques such as CRISPR/Cas9 (Jansing et al., 2018). Transgenic *N. benthamiana* lines that could be used systematically to produce recombinant proteins were therefore generated in many research labs, including our university (Strasser et al., 2008).

The transgenic *N. benthamiana* ΔXT/FT line (Strasser et al., 2008) was generated to handle the production *in planta* of human-like recombinant glycoproteins. Producing recombinant proteins *in planta* requires special handling of the plant's glycosylation machinery (Montero-Morales & Steinkellner, 2018), since plants and humans have different glycosylation patterns. This is particularly relevant if the recombinant protein produced has a therapeutic usage: in fact, a wrong glycosylation might alter the fold or the activity of the recombinant protein (Shental-Bechor & Levy, 2008; Skropeta, 2009), and the potential antigenicity to humans of plant-specific glycans is still under debate (Bosch, Castilho, Loos, Schots, & Steinkellner, 2013; Dowling et al., 2007; Lisowska, 2002; Rup et al., 2017). A correlation between antibody potency and core fucosylation has also been mentioned (Alexandra Castilho et al., 2015).

The genome and the transcriptome of ΔXT/FT were thoroughly characterized in this work (Schiavinato et al., 2019). This line specifically knocks down the expression of fucosyl-transferases (FucT) and xylosyl-transferases (XylT), preventing the addition of the fucose and the xylose glycan residues to the growing glycan chains of newly produced proteins (Strasser et al., 2008). This plant has been successfully used to produce antibodies in many studies (A Castilho et al., 2011; Dent et al., 2016; Montero-Morales et al., 2017), highlighting the relevance of *N. benthamiana* in this field of research. In the first attached publication (Schiavinato et al., 2019), the details of the ΔXT/FT characterization are reported.

This project

This thesis describes a series of analyses aimed at furthering the knowledge on *N. benthamiana*'s genome and transcriptome. As mentioned in the background chapter, a large number of *N. benthamiana* mRNA-Seq data points (2.35 billion mRNA-Seq reads) accumulated in public repositories such as SRA (Leinonen, Sugawara, Shumway, & International Nucleotide Sequence Database Collaboration, 2011) and Sol Genomics Network (Fernandez-Pozo et al., 2015). These reads were obtained from a variety of tissues and conditions (Figure **7**). Of these, 126 million reads were generated by us.

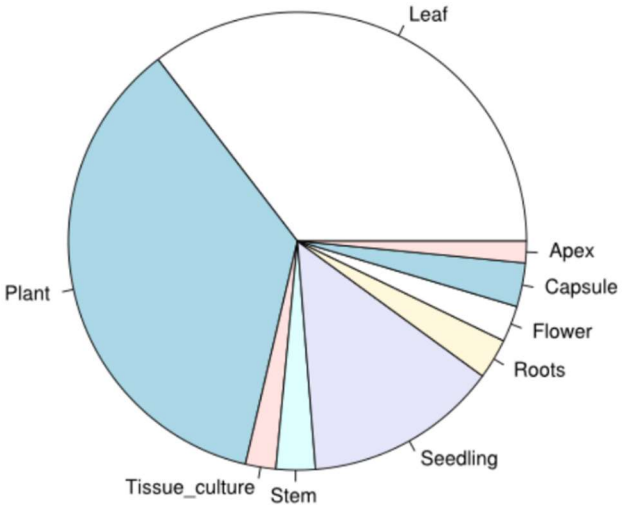


Figure 7: mRNASeq reads by tissue. Each slice represents a different tissue from which mRNASeq reads were taken. "Plant" refers to unspecified tissues. The abundance is quantified in number of reads.

With this data in hand, I performed a robust gene prediction on the Nb-1 draft genome assembly (Bombarely et al., 2012). The gene prediction was realized using state-of-the-art *in silico* methods that could harness the full potential of mRNA-Seq information (Figure **8**). The produced gene set, referred to as "NibSet-1", has been used throughout the entire thesis and represented the first milestone of this work. Making use of homology information with other plant species (mostly *Solanaceae*), I produced a functional annotation for it. The final product was a supported, annotated and validated gene set of 62,216 transcripts organized in 50,516 genes. This gene set was uploaded to a public repository that is linked in the first attached publication (Schiavinato et al., 2019).

The first application of this gene set has been the transcriptional profiling of the ΔXT/FT *N. benthamiana* line (Strasser et al., 2008). As described before, this line operates a knock-down of the FucT and XylT genes in order to produce human recombinant glycoproteins *in planta*. The aim of this part of the project was to understand whether the knock-down of these genes,

operated via two transgenes inserted in the genome, was affecting the transcription of any other gene. Together with the differentially expressed genes (DEGs), we also characterized the transgene insertion sites in the genome. We observed one of the transgenes inserted within an active gene. The fusion of the two open reading frames produced a chimeric transcript that was among the upregulated genes found. The details of this part of the work are also contained in Schiavinato et al. (2019).

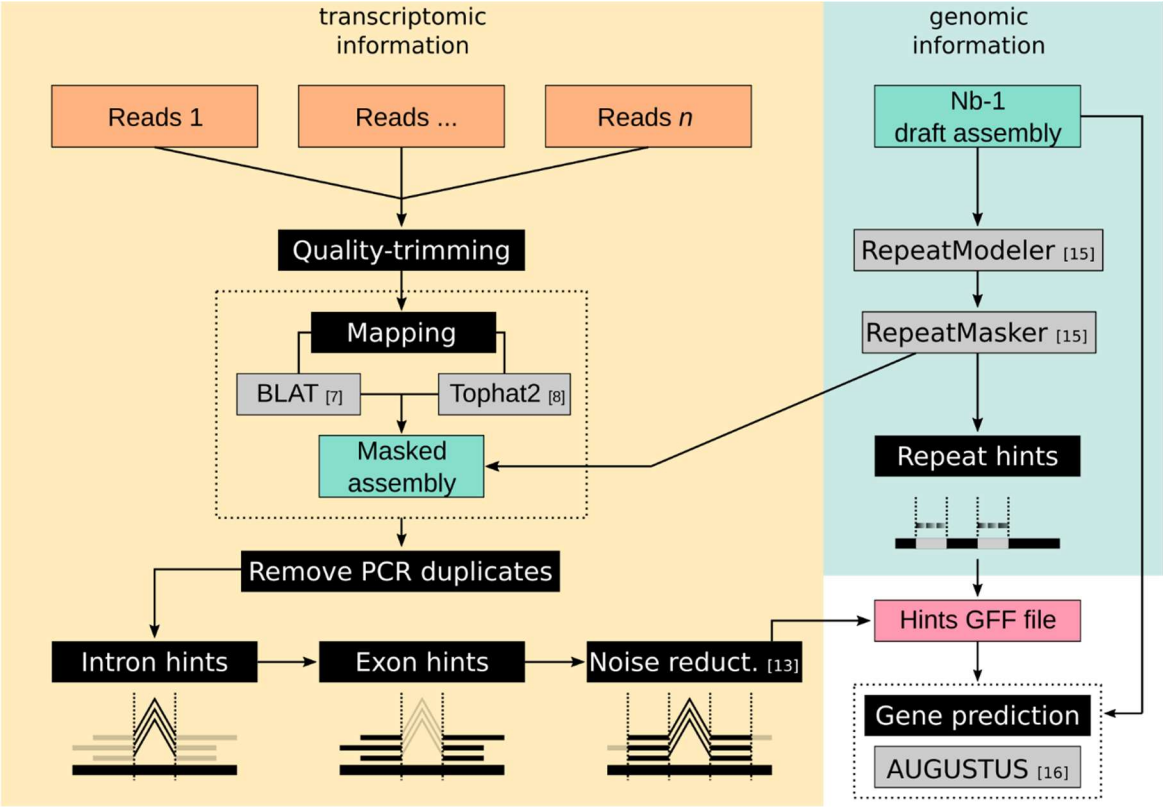


Figure 8: Gene prediction workflow. The orange box describes how the expression information, deriving from mRNASeq reads, is obtained. The light green box shows how the information on repetitive regions is integrated. BLAT **(Kent, 2002)**, Tophat2 **(Kim et al., 2013)**, RepeatModeler **(Smit & Hubley, 2008)**, RepeatMasker **(Smit et al., 2013)**, Augustus **(Stanke et al., 2006)** and the noise reduction Perl script **(Minoche et al., 2015)** are the softwares used to achieve the gene prediction.

Another point that was addressed in the first attached publication is the genome relatedness between *N. benthamiana* research lines. Many research lines used around the world lack proper documentation and often it is not known where they were taken from. A bright example of this can be seen in the two major reviews published in the last twelve years (Bally et al., 2018; Goodin et al., 2008). In the most recent of the two, it is described how new documents were discovered (dated to the 1930s) which could finally trace how *N. benthamiana* arrived in the US research centres. The older review, instead, still had this as an open question. The main question has been, in fact, whether each research accession used in the world derived from the same original plant specimen, or whether they originate from multiple plants collected

independently. To address this issue, I studied line-specific variations using sequencing data mapped on the Nb-1 draft genome assembly.

We then turned our attention to *N. benthamiana*'s evolutionary history. Prior to this work, the knowledge on this topic was limited. As discussed before, the parental progenitors were addressed multiple times, but no strong conclusion was taken on the maternal progenitor (Kelly et al., 2012). We attributed this to a general lack of data, which limited the analyses to a small set of genes. We also discussed how the hybridization date was estimated differently in multiple studies (Chase, 2003; Clarkson et al., 2017; Leitch et al., 2008). Both questions could be answered by making use of the NibSet-1 gene set together with gene sets from other diploid *Nicotiana* species. We detected the maternal progenitor using homology relationships of NibSet-1 genes with several diploid species of the *Nicotiana* genus, generating thousands of phylogenetic trees, collectively referred to as a phylome (Huerta-Cepas, Capella-Gutiérrez, Pryszcz, Marcet-Houben, & Gabaldón, 2014). Contrary to previous studies, which were based either on hand-picked gene trees or on few hundreds of reliable ones, the use of a phylome allowed us to have a more weighted and general overview, with thousands of trees evaluated at once. With a series of filters, I extracted the most reliable phylogenetic trees and counted the occurrences of each candidate parental species (Figure **9**, taken from Schiavinato et al, in press).

The phylome was also used to study the hybridization date. Such a large collection of phylogenetic trees could, in fact, be used to estimated divergence times between *N. benthamiana* and each candidate parent. To address this, I made use of fourfold-degenerate (4D) codon sites, which are neutrally evolving positions (Lagerkvist, 1978). Being almost free of any selective pressure, they represent a good footprint of time. I first studied the transversion rate at 4D sites (4DTv), which is a relative measurement of time within a data set. Transversions accumulate slowly, hence a larger proportion of transversions implies that more time has passed. In each gene tree of the phylome I computed the 4DTv ratio between *N. benthamiana* and the other diploid species. I then studied the silent substitutions at 4D sites between *N. benthamiana* genes and their homologs in diploid species, regardless of the nature of the substitution (i.e. both transitions and transversions).

I compared the number of silent substitutions to the expected mutation rate of plant nuclear genes (Wolfe, Li, & Sharp, 1987), and then converted this value to millions of years. I then used the distribution of the values to obtain an estimation of the most likely hybridization date, comparing distributions of the two candidate parental species found with the phylome. The details are contained in the second attached publication (Schiavinato et al, in press).

In the same publication we studied in more detail the status of the two subgenomes contained within the *N. benthamiana* genome. Specifically, we aimed at separating the two subgenomes to understand the extent of LTGD and subgenomic intermixing. In fact, the subgenomes of *N.*

benthamiana are highly intermixed due to its high genome plasticity that allows homoeologous recombination and genomic rearrangements more easily than other species (Bally et al., 2015).

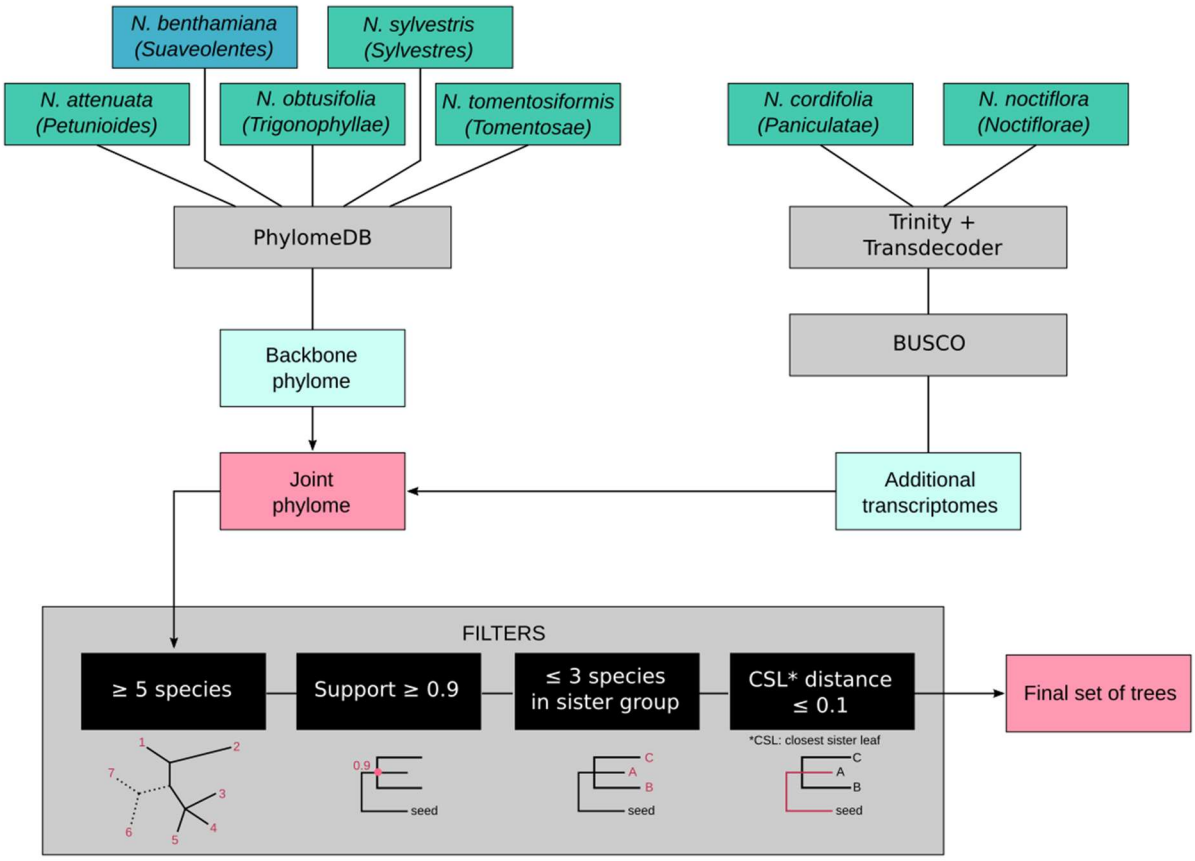


Figure 9: Phylome analysis workflow. Workflow used in the generation of the *N. benthamiana* phylome. The backbone phylome was generated from gene models annotated within the sequenced genomes of five species of tobacco, including *N. benthamiana* as the seed species (blue box). To the backbone phylome, two more species (*N. cordifolia*, *N. noctiflora*) were added based on genes obtained from transcriptome assembly. Sections are indicated below the species names. Software and procedures are included in grey boxes. Red boxes indicate phylogenetic tree collections.

Subgenome separation is traditionally performed with read mapping. Reads from the two parental progenitors are generated and mapped onto the hybrid reference genome sequence. Then, each chromosome is assigned to a parental progenitor depending on which set of reads represented it better (i.e. generated a better coverage profile). This rationale, however, holds only until the two subgenomes are sufficiently separate. In the specific case of *N. benthamiana*, its genome plasticity made it difficult to assign entire chromosomes to a subgenome. We could, however, retrieve the parental origin of genes from the phylome trees, and use it as a proxy to assign fractions of the genome to either one of the subgenomes (Schiavinato et al, in press).

References

Anderson, E., & Stebbins, G. L. (1954). Hybridization as an evolutionary stimulus. *Evolution*, *8*, 378–388.

Aston, C., Mishra, B., & Schwartz, D. C. (1999). Optical mapping and its potential for large-scale sequencing projects. *Trends in Biotechnology*, *17*, 297–302.

Bally, J., Jung, H., Mortimer, C., Naim, F., Philips, J. G., Hellens, R., … Waterhouse, P. M. (2018). The Rise and Rise of *Nicotiana benthamiana*: A Plant for All Reasons. *Annual Review of Phytopathology*, *56*, 405–426.

Bally, J., Nakasugi, K., Jia, F., Jung, H., Ho, S. Y. W., Wong, M., … Waterhouse, P. M. (2015). The extremophile *Nicotiana benthamiana* has traded viral defence for early vigour. *Nature Plants*, *1*, 15165.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *J.L. Plant Mol Biol*, *42*, 251–269.

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., … Schmutz, J. (2019). The genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nature Genetics*, *51*, 877–884.

Biscotti, M. A., Olmo, E., & Heslop-Harrison, J. S. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome Research*, *23*, 415–420.

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C. S., … Kuhlemeier, C. (2016). Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. *Nature Plants*, *2*, 16074.

Bombarely, A., Rosli, H. G., Vrebalov, J., Moffett, P., Mueller, L. A., & Martin, G. B. (2012). A Draft Genome Sequence of *Nicotiana benthamiana* to Enhance Molecular Plant-Microbe Biology Research. *Molecular Plant-Microbe Interactions*, *25*, 1523–1530.

Bosch, D., Castilho, A., Loos, A., Schots, A., & Steinkellner, H. (2013). N-glycosylation of plant-produced recombinant proteins. *Current Pharmaceutical Design*, *19*, 5503–5512.

Canady, M. A., Ji, Y., & Chetelat, R. T. (2006). Homeologous Recombination in *Solanum lycopersicoides* Introgression Lines of Cultivated Tomato. *Genetics*, *174*, 1775–1788.

Castilho, A, Bohorova, N., Grass, J., Bohorov, O., Zeitlin, L., Whaley, K., … Steinkellner, H. (2011). Rapid High Yield Production of Different Glycoforms of Ebola Virus Monoclonal Antibody. *PLoS ONE*, *6*, e26040.

Castilho, Alexandra, Gruber, C., Thader, A., Oostenbrink, C., Pechlaner, M., Steinkellner, H., & Altmann, F. (2015). Processing of complex N-glycans in IgG Fc-region is affected by core fucosylation. *MAbs*, *7*, 863–870.

Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2008). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*, *19*, 336–346.

Chase, M. W. (2003). Molecular Systematics, GISH and the Origin of Hybrid Taxa in Nicotiana (Solanaceae). *Annals of Botany*, *92*, 107–127.

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., & Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nature Plants*, *4*, 258–268.

Clarkson, J. J., Dodsworth, S., & Chase, M. W. (2017). Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in *Nicotiana* (*Solanaceae*). *Plant Systematics and Evolution*, *303*, 1001–1012.

Clarkson, J. J., Lim, K. Y., Kovarik, A., Chase, M. W., Knapp, S., & Leitch, A. R. (2005). Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (*Solanaceae*). *New Phytologist*, *168*, 241–252.

Conley, A. J., Zhu, H., Le, L. C., Jevnikar, A. M., Lee, B. H., Brandle, J. E., & Menassa, R. (2011). Recombinant protein production in a variety of Nicotiana hosts: A comparative analysis: Protein production in various Nicotiana hosts. *Plant Biotechnology Journal*, *9*, 434–444.

Dent, M., Hurtado, J., Paul, A. M., Sun, H., Lai, H., Yang, M., … Chen, Q. (2016). Plant-produced anti-dengue virus monoclonal antibodies exhibit reduced antibody-dependent enhancement of infection activity. *Journal of General Virology*, *97*, 3280–3290.

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*, e105.

Dohm, J. C., Minoche, A. E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., … Himmelbauer, H. (2013). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, *505*, 546–549.

Dowling, W., Thompson, E., Badger, C., Mellquist, J. L., Garrison, A. R., Smith, J. M., … Schmaljohn, C. (2007). Influences of Glycosylation on Antigenicity, Immunogenicity, and Protective Efficacy of Ebola Virus GP DNA Vaccines. *Journal of Virology*, *81*, 1821–1837.

Edwards, K. D., Fernandez-Pozo, N., Drake-Stowe, K., Humphry, M., Evans, A. D., Bombarely, A., … Mueller, L. A. (2017). A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics*, *18*, 448.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, *323*, 133–138.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., … Mueller, L. A. (2015). The Sol Genomics Network (SGN)—From genotype to phenotype to breeding. *Nucleic Acids Research*, *43*, D1036–D1041.

Fischer, R., & Emans, N. (2000). Molecular farming of pharmaceutical proteins. *Transgenic Research*, *9*, 279–299; discussion 277.

Flavell, A. J., Pearce, S. R., & Kumar, A. (1994). Plant transposable elements and the genome. *Current Opinion in Genetics & Development*, *4*, 838–844.

Glover, N. M., Redestig, H., & Dessimoz, C. (2016). Homoeologs: What Are They and How Do We Infer Them? *Trends in Plant Science*, *21*, 609–621.

Goodin, M. M., Zaitlin, D., Naidu, R. A., & Lommel, S. A. (2008). *Nicotiana benthamiana*: Its History and Future as a Model for Plant–Pathogen Interactions. *Molecular Plant-Microbe Interactions*, *2008*, 28–39.

Goodspeed, T. (1954). The genus nicotiana. By Thomas Harper Goodspeed. Chronica Botanica Company, Waltham, Mass., 1954, and Stechert-Hafner, Inc., New York, 1955. Illustrated. Xxii+536 pp. 16.5 × 25 cm. Price $12.50. *Journal of the American Pharmaceutical Association (Scientific Ed.)*, *45*, 193.

Grandbastien, M.-A., Spielmann, A., & Caboche, M. (1989). Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature*, *337*, 376–380.

Grant, V. (1981). *Plant speciation*. Columbia University Press,.

Hanson, R. E., Zhao, X., Islam-Faridi, M. N., Paterson, A. H., Zwick, M. S., Crane, C. F., … Price, H. J. (1998). Evolution of interspersed repetitive elements in *Gossypium* (Malvaceae). *American Journal of Botany*, *85*, 1364–1368.

Harrison, R. G., & others. (1990). Hybrid zones: Windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, *7*, 69–128.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M., & Gabaldón, T. (2014). PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Research*, *42*, D897–D902.

Imelfort, M., & Edwards, D. (2009). De novo sequencing of plant genomes using second-generation technologies. *Briefings in Bioinformatics*, *10*, 609–618.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*, 239.

Jansing, J., Sack, M., Augustine, S. M., Fischer, R., & Bortesi, L. (2018). CRISPR/Cas9-mediated knockout of six glycosyltransferase genes in *Nicotiana benthamiana* for the production of recombinant proteins lacking β-1,2-xylose and core α-1,3-fucose. *Plant Biotechnology Journal*. https://doi.org/10.1111/pbi.12981

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., … Tester, M. (2017). The genome of Chenopodium quinoa. *Nature*, *542*, 307–312.

Joensuu, J. J., Conley, A. J., Lienemann, M., Brandle, J. E., Linder, M. B., & Menassa, R. (2010). Hydrophobin Fusions for High-Level Transient Protein Expression and Purification in Nicotiana benthamiana. *PLANT PHYSIOLOGY*, *152*, 622–633.

Julca, I., Marcet-Houben, M., Vargas, P., & Gabaldón, T. (2018). Phylogenomics of the olive tree (Olea europaea) reveals the relative contribution of ancient allo- and autopolyploidization events. *BMC Biology*, *16*, 15.

Kejnovsky, E., Hawkins, J. S., & Feschotte, C. (2012). Plant Transposable Elements: Biology and Evolution. In J. F. Wendel, J. Greilhuber, J. Dolezel, & I. J. Leitch (Eds.), *Plant Genome Diversity Volume 1* (pp. 17–34). Vienna: Springer Vienna.

Kelly, L. J., Leitch, A. R., Clarkson, J. J., Knapp, S., & Chase, M. W. (2012). Reconstructing the Complex Evolutionary Origin of Wild Allopolyploid Tobaccos (*Nicotiana* section *Suaveolentes*). *Evolution*, *67*, 80–94.

Kent, W. J. (2002). BLAT---The BLAST-Like Alignment Tool. *Genome Research*, *12*, 656–664.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*, 357–360.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*, R36.

Knapp, S., Chase, M. W., & Clarkson, J. J. (2004). Nomenclatural Changes and a New Sectional Classification in *Nicotiana* (*Solanaceae*). *Taxon*, *53*, 73.

Kościańska, E., Kalantidis, K., Wypijewski, K., Sadowski, J., & Tabler, M. (2005). Analysis of RNA Silencing in Agroinfiltrated Leaves of Nicotiana Benthamiana and Nicotiana Tabacum. *Plant Molecular Biology*, *59*, 647–661.

Lagerkvist, U. (1978). "Two out of three": An alternative method for codon reading. *Proceedings of the National Academy of Sciences of the United States of America*, *75*, 1759–1762.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359.

Laudencia-Chingcuanco, D., & Fowler, D. B. (2012). Genotype-dependent burst of transposable element expression in crowns of hexaploid wheat (Triticum aestivum L.) during cold acclimation. *Comparative and Functional Genomics*, *2012*.

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*, D19-21.

Leitch, I. J., Hanson, L., Lim, K. Y., Kovarik, A., Chase, M. W., Clarkson, J. J., & Leitch, A. R. (2008). The Ups and Downs of Genome Size Evolution in Polyploid Species of *Nicotiana* (*Solanaceae*). *Annals of Botany*, *101*, 805–814.

Levin, D. A. (2000). *The origin, expansion, and demise of plant species*. Oxford University Press on Demand.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., … Yu, S. (2015). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature Biotechnology*, *33*, 524–530.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *26*, 589–595.

Li, J., Stoddard, T. J., Demorest, Z. L., Lavoie, P.-O., Luo, S., Clasen, B. M., … Zhang, F. (2016). Multiplexed, targeted gene editing in *Nicotiana benthamiana* for glyco-engineering and monoclonal antibody production. *Plant Biotechnology Journal*, *14*, 533–542.

Lim, K. Yoong, Kovarik, A., Matyasek, R., Chase, M. W., Clarkson, J. J., Grandbastien, M. A., & Leitch, A. R. (2007). Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytologist*, *175*, 756–763.

Lim, K.Y., Matyasek, R., Kovarik, A., Fulnecek, J., & Leitch, A. R. (2005). Molecular cytogenetics and tandem repeat sequence evolution in the allopolyploid *Nicotiana rustica* compared with diploid progenitors *N. paniculata* and *N. undulata*. *Cytogenetic and Genome Research*, *109*, 298–309.

Lisowska, E. (2002). The role of glycosylation in protein antigenic properties. *Cellular and Molecular Life Sciences: CMLS*, *59*, 445–455.

Liu, L., Zhang, Y., Tang, S., Zhao, Q., Zhang, Z., Zhang, H., … Xie, Q. (2010). An efficient system to detect protein ubiquitination by agroinfiltration in *Nicotiana benthamiana*. *The Plant Journal*, *61*, 893–903.

Long, N., Ren, X., Xiang, Z., Wan, W., & Dong, Y. (2016). Sequencing and characterization of leaf transcriptomes of six diploid *Nicotiana* species. *Journal of Biological Research-Thessaloniki*, *23*, 6.

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*, 265–279.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*. https://doi.org/10.1186/2047-217X-1-18

Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., … Terauchi, R. (2003). Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 15718–15723.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, *226*, 792–801.

Meyers, B. C. (2001). Abundance, Distribution, and Transcriptional Activity of Repetitive Elements in the Maize Genome. *Genome Research*, *11*, 1660–1676.

Meyers, Blake C., Scalabrin, S., & Morgante, M. (2004). Mapping and sequencing complex genomes: Let's get physical! *Nature Reviews Genetics*, *5*, 578–588.

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, *12*, R112.

Minoche, A. E., Dohm, J. C., Schneider, J., Holtgräwe, D., Viehöver, P., Montfort, M., … Himmelbauer, H. (2015). Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biology*, *16*, 184.

Montero-Morales, L., Maresch, D., Castilho, A., Turupcu, A., Ilieva, K. M., Crescioli, S., … Steinkellner, H. (2017). Recombinant plant-derived human IgE glycoproteomics. *Journal of Proteomics*, *161*, 81–87.

Montero-Morales, L., & Steinkellner, H. (2018). Advanced Plant-Based Glycan Engineering. *Frontiers in Bioengineering and Biotechnology*, *6*, 81.

Naim, F., Nakasugi, K., Crowhurst, R. N., Hilario, E., Zwart, A. B., Hellens, R. P., … Wood, C. C. (2012). Advanced Engineering of Lipid Metabolism in *Nicotiana benthamiana* Using a Draft Genome and the V2 Viral Silencing-Suppressor Protein. *PLoS ONE*, *7*, e52717.

Nakasugi, K., Crowhurst, R., Bally, J., & Waterhouse, P. (2014). Combining Transcriptome Assemblies from Multiple De Novo Assemblers in the Allo-Tetraploid Plant *Nicotiana benthamiana*. *PLoS ONE*, *9*, e91776.

Petit, M., Guidat, C., Daniel, J., Denis, E., Montoriol, E., Bui, Q. T., … Mhiri, C. (2010). Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytologist*, *186*, 135–147.

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, *13*, 341.

Renny-Byfield, S., Gong, L., Gallagher, J. P., & Wendel, J. F. (2015). Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution. *Molecular Biology and Evolution*, *32*, 1063–1071.

Renny-Byfield, S., Kovarik, A., Kelly, L. J., Macas, J., Novak, P., Chase, M. W., … Leitch, A. R. (2013). Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *The Plant Journal*, *74*, 829–839.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., … Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*, R51.

Rup, B., Alon, S., Amit-Cohen, B.-C., Brill Almon, E., Chertkoff, R., Tekoah, Y., & Rudd, P. M. (2017). Immunogenicity of glycans on biotherapeutic drugs produced in plant expression systems-The taliglucerase alfa story. *PloS One*, *12*, e0186211.

Schiavinato, M., Strasser, R., Mach, L., Dohm, J. C., & Himmelbauer, H. (2019). Genome and transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔXT/FT. *BMC Genomics*, *20*, 594.

Schillberg, S., Emans, N., & Fischer, R. (2002). Antibody molecular farming in plants and plant cells. *Phytochemistry Reviews*, *1*, 45–54.

Shental-Bechor, D., & Levy, Y. (2008). Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 8256–8261.

Sierro, N., Battey, J. N. D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., … Ivanov, N. V. (2014). The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications*, *5*, 3833.

Skropeta, D. (2009). The effect of individual N-glycans on enzyme activity. *Bioorganic & Medicinal Chemistry*, *17*, 2645–2653.

Smit, A., & Hubley, R. (2008). *RepeatModeler*. Retrieved from http://www.repeatmasker.org

Smit, A., Hubley, R., & Green, P. (2013). *RepeatMasker*. Retrieved from http://www.repeatmasker.org

Soltis, P. S., Marchant, D. B., Van de Peer, Y., & Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development*, *35*, 119–125.

Soltis, P. S., & Soltis, D. E. (2009). The Role of Hybridization in Plant Speciation. *Annual Review of Plant Biology*, *60*, 561–588.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: *Ab initio* prediction of alternative transcripts. *Nucleic Acids Research*, *34*, W435-439.

Strasser, R., Stadlmann, J., Schähs, M., Stiegler, G., Quendler, H., Mach, L., … Steinkellner, H. (2008). Generation of glyco-engineered *Nicotiana benthamiana* for the production of monoclonal antibodies with a homogeneous human-like N-glycan structure. *Plant Biotechnology Journal*, *6*, 392–402.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, *13*, 36–46.

Udall, J. A., Quijada, P. A., & Osborn, T. C. (2005). Detection of Chromosomal Rearrangements Derived From Homeologous Recombination in Four Mapping Populations of *Brassica napus* L. *Genetics*, *169*, 967–979.

van Herpen, T. W. J. M., Cankar, K., Nogueira, M., Bosch, D., Bouwmeester, H. J., & Beekwilder, J. (2010). *Nicotiana benthamiana* as a Production Platform for Artemisinin Precursors. *PLoS ONE*, *5*, e14222.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., … Viola, R. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). *Nature Genetics*, *42*, 833–839.

Vision, T. J., Brown, D. G., & Tanksley, S. D. (2000). The Origins of Genomic Duplications in Arabidopsis. *Science*, *290*, 2114–2117.

Wiley, E. O. (1978). The evolutionary species concept reconsidered. *Systematic Zoology*, *27*, 17–26.

Wolfe, K. H., Li, W. H., & Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America*, *84*, 9054–9058.

Wydro, M., Kozubek, E., & Lehmann, P. (2006). Optimization of transient Agrobacterium-mediated gene expression system in leaves of Nicotiana benthamiana. *Acta Biochimica Polonica*, *53*, 289–298.

Chapter 2:

Genome and transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔXT/FT

Published as the article:

Schiavinato, M., Strasser, R., Mach, L., Dohm, J.C. and Himmelbauer, H., Genome and transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔXT/FT. *BMC Genomics* 20, 594 (2019) doi:10.1186/s12864-019-5960-2

Schiavinato et al. BMC Genomics (2019) 20:594
https://doi.org/10.1186/s12864-019-5960-2

BMC Genomics

RESEARCH ARTICLE Open Access

Genome and transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔXT/FT

Matteo Schiavinato[1], Richard Strasser[2], Lukas Mach[2], Juliane C. Dohm[1*] and Heinz Himmelbauer[1*]

Abstract

Background: The allotetraploid tobacco species *Nicotiana benthamiana* native to Australia has become a popular host for recombinant protein production. Although its usage grows every year, little is known on this plant's genomic and transcriptomic features. Most *N. benthamiana* accessions currently used in research lack proper documentation of their breeding history and provenance. One of these, the glycoengineered *N. benthamiana* line ΔXT/FT is increasingly used for the production of biopharmaceutical proteins.

Results: Based on an existing draft assembly of the *N. benthamiana* genome we predict 50,516 protein –encoding genes (62,216 transcripts) supported by expression data derived from 2.35 billion mRNA-seq reads. Using single-copy core genes we show high completeness of the predicted gene set. We functionally annotate more than two thirds of the gene set through sequence homology to genes from other *Nicotiana* species. We demonstrate that the expression profiles from leaf tissue of ΔXT/FT and its wild type progenitor only show minimal differences. We identify the transgene insertion sites in ΔXT/FT and show that one of the transgenes was inserted inside another predicted gene that most likely lost its function upon insertion. Based on publicly available mRNA-seq data, we confirm that the *N. benthamiana* accessions used by different research institutions most likely derive from a single source.

Conclusions: This work provides gene annotation of the *N. benthamiana* genome, a genomic and transcriptomic characterization of a transgenic *N. benthamiana* line in comparison to its wild-type progenitor, and sheds light onto the relatedness of *N. benthamiana* accessions that are used in laboratories around the world.

Keywords: *Nicotiana benthamiana*, Genome, Gene prediction, Transgene, Intraspecific variation, Accession history

Background

Nicotiana benthamiana is an allotetraploid plant indigenous to Australia. The *Nicotiana* genus is a member of the *Solanaceae* family which is particularly relevant in agriculture, and includes potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*), eggplant (*Solanum melongena*), and the smoking tobacco (*Nicotiana tabacum*). The fame of *N. benthamiana* is however mostly due to its versatility for studies of plant-pathogen interaction and molecular farming rather than crop sciences [1–4]. During the last two decades this plant emerged as a very promising host for recombinant protein production, in particular for medical application as vaccines or antibodies [5–7].

Most prominently, the transgenic *N. benthamiana* line ΔXT/FT has been engineered [8] to act as a production system for therapeutic proteins and has been successfully used to produce antibodies at an industrial scale [5, 9, 10]. Its main feature is the knockdown of genes encoding fucosyl-transferases (FT) and xylosyl-transferases (XT) through RNA interference, a procedure that enables the production of recombinant glycoproteins with human glycan profiles *in planta*. Glycans influence protein folding and modulate protein activity [11, 12], and there is evidence that plant-specific glycan structures could potentially be antigenic to humans [13–15], even though this has been recently debated [16]. A linkage between core fucosylation and monoclonal antibody potency has also been described [17].

* Correspondence: dohm@boku.ac.at; heinz.himmelbauer@boku.ac.at
[1]Department of Biotechnology, University of Natural Resources and Life Sciences (BOKU), Muthgasse 18, 1190 Vienna, Austria
Full list of author information is available at the end of the article

Despite *N. benthamiana*'s widespread use in research, and its growing importance as an expression platform for recombinant proteins, comparatively little is known about its genomic and transcriptomic features on the sequence level. In 2012, a first milestone was achieved with the publication of the Nb-1 draft genome assembly [18] that is available at the SOL Genomics Network website (https://solgenomics.net/) [19]. This assembly covers around 86% of the haploid genome size of *N. benthamiana*, which is estimated at 3.136 Gbp [3]. Another draft genome assembly was published the same year from a different research group [20], which published also a de novo assembled transcriptome in the following years [21, 22]. We also note the publication of a recent *N. benthamiana* gene set, which was obtained from mapping of genes identified in other *Nicotiana* species onto the *N. benthamiana* genome [23]. Here, we perform evidence-based gene prediction supported by 2.35 billion mRNA-seq reads and characterize the transcriptome. We use our predicted gene set to carry out genomic and transcriptomic analyses of the glycoengineered *N. benthamiana* line ΔXT/FT. We address the question where the two RNA interference cassettes have been inserted within the genome, and if the insertions might impact gene expression. For these comparisons, we generated additional high-coverage genomic and transcriptomic datasets from our parental *N. benthamiana* wild type line (WT) as well as the glycoengineered line ΔXT/FT derived thereof. We use transcriptomic data to explore whole-transcriptome differential expression between ΔXT/FT and WT, and we use the genomic data to identify single-nucleotide variants (SNVs) and insertion/deletion variants (indels) and discuss their functional impact. Finally, we address inter-accession relatedness between *N. benthamiana* lines in use at different research institutions. The lack of documentation for most of these lines makes it challenging to understand their real genetic diversity. The reproducibility of experimental results could in fact depend heavily on the genotype of the accession. By assessing the variants found within annotated coding regions of the *N. benthamiana* genome, we attempt to characterize this diversity.

Results

N. benthamiana gene catalogue and functional annotation

The Nb-1 draft genome assembly [18] comprising a total size of 2.97 Gbp with an N50 size of 0.5 Mbp was used as starting point to predict a gene set for *N. benthamiana*. We identified 60.7% of the sequence (excluding Ns) being composed of transposable elements (TEs) of which the majority belonged to the class of LTR retrotransposons (Additional file 1: Table S1), as expected for plant genomes [24, 25]. On the TE-masked Nb-1 genome we performed gene prediction using the Augustus pipeline [26]. A particular strength of Augustus is its combination of in silico gene prediction and integration of evidence from transcriptome sequencing, providing experimental support for the predictions. As transcriptomic evidence a total of 2.35 billion mRNA-seq reads from eight different *N. benthamiana* accessions were used, corresponding to 151.6 Gb of sequencing data; of these, 126 million reads (31.5 Gb) were generated in this study (Additional file 2). Data sources were chosen in a way that multiple tissues and stress conditions were represented. From 114,605 initial predictions we kept 62,216 transcripts (50,516 genes) that were supported by at least 1% mRNA-seq evidence and had no major overlap (max. 10 nt) with annotated TEs in coding regions; thirteen peptides of less than ten amino acids were removed from the set of protein sequences. The final set of gene predictions is referred to as "NibSet-1". The average gene length including introns was 5,573 bp, the average transcript length was 1,665 bp, and the average protein length was 404 amino acids. The average number of exons per transcript was 6.2, and 59,410 transcript models (95.5%) included both start and stop codon (Table 1). Notably, 30,974 (61.3%) of the predicted gene models were fully supported by mRNA-seq evidence, i.e. all their predicted features, such as exon-intron junctions and UTRs, were supported by transcriptomic reads.

We used the fully supported models to test if they extend the gene set of an older gene prediction available at the SOL Genomics Network website [19], called Niben101_annotation. Most of the NibSet-1 high-confidence genes (26,817 of 30,974; 86.6%) overlapped at least for half of their length with a Niben101_annotation model of which 6,364 coincided perfectly when comparing

Table 1 *N. benthamiana* NibSet-1 gene set metrics

Genes	50,516
Transcripts	62,216
Protein sequences	62,203
Multi-isoform genes	8,676
Transcripts with start and stop codons	59,410
Average gene length	5,573 nt
Average transcript length	1,665 nt
Average number of exons per transcript	6.2
Number of single-exon transcripts	7,410
Average exon length	268 nt
Average length of coding exon (CDS)	213 nt
Average intron length	801 nt
Average protein length	404 aa

annotated CDS coordinates. To verify the remaining 4,157 high-confidence NibSet-1 gene models we mapped them against the transcriptome of the paternal progenitor *Nicotiana sylvestris*. A large fraction (3,651 genes, 87.8%) found a match in *N. sylvestris* (minimum 90% sequence identity) and, hence, are likely to represent true genes that were missing in Niben101_annotation. We concluded that given the high amount of mRNA-seq data supporting our gene models, NibSet-1 is likely to be more accurate than Niben101_annotation and that NibSet-1 provides additional high-confidence genes that complement the gene models of Niben101_annotation. We also noted that the average protein length of Niben101_annotation was smaller (327 amino acids) than in NibSet-1 (404 amino acids, see above), suggesting that NibSet-1 was less fragmented than Niben101_annotation.

We validated the completeness of NibSet-1 by searching for sequence homology in a set of highly conserved plant genes using BUSCO (benchmarking universal single-copy orthologs) [27]. Out of 956 conserved plant genes, 937 (98.0%) were matched by a predicted *N. benthamiana* sequence (only one transcript per gene was used). For the sake of comparison, we ran BUSCO also on the Niben101_annotation gene set: 932 (97.5%) conserved plant genes were found (Additional file 1: Table S2) showing that highly conserved genes are well represented in both gene sets with a slightly higher level of completeness in NibSet-1 compared to Niben101_annotation.

Public NCBI databases [28] contained 401 *N. benthamiana* protein sequences (as of June 2017), of which 396 (98.8%) matched NibSet-1 protein sequences with a minimum sequence identity of 95%. All 401 sequences found a match with ≥85% sequence identity. Overall, we consider NibSet-1 to be a highly complete and accurate representation of *N. benthamiana*'s gene repertoire.

We functionally annotated the NibSet-1 protein sequences by transferring annotations from homologous genes of other plant species (Additional file 1: Table S3) with sequence similarity ≥ 90% and alignment length ≥ 70 amino acids. In total, we assigned functional annotations to 44,184 (71%) *N. benthamiana* protein sequences belonging to 35,428 genes (Fig. 1). The majority (42,344 proteins, 95.8%) was annotated through homologous sequences from the *Nicotiana* genus, further annotations were transferred from the *Solanaceae* family (27 proteins), *Arabidopsis* (13 proteins), and "non-redundant" NCBI databases (1,800 proteins). Only 1,549 (2.5%) protein sequences corresponding to 1,499 genes could not find a match in any of the tested datasets.

Characterization of transgene integration sites in the *N. benthamiana* line ΔXT/FT

The glycoengineered ΔXT/FT *N. benthamiana* line was generated to avoid the addition of the plant-specific glycan residues β1,2-xylose and core α1,3-fucose to recombinantly produced glycoproteins. This was achieved via the insertion of two transgenes (Additional file 3), which mediate down-regulation of the genes encoding core α1, 3-fucosyltransferase (FucT) and β1,2-xylosyltransferase (XylT) by means of RNA interference [8]. In a recent study, five FucT genes have been described, with one of them probably representing a pseudogene [29]. Our raw gene set, prior to any filtering step, included all of them, i.e. FucT1 = g31184, FucT2 = g80352, FucT3 = g3481, FucT4 = g97519, FucT5 = g36277; gene g97519 was later removed due to an overlap with annotated transposable elements. The transgenes used in the glycoengineered ΔXT/FT *N. benthamiana* line were designed to act on at least two FucT genes (g31184 and g80352 in NibSet-1) and on both XylT genes (g40438 and g43728). We replaced Augustus FucT and XylT gene models in NibSet-1 (g31184, g40438, g43728, g80352) with the corresponding manually curated sequences from Strasser et al. (2008) (sequence identity 99%, see Additional file 1: Text; Figure S1).

Transgene insertion into the host genome occurs at positions that cannot be predicted [30]; it is therefore important to assess potential unintended changes to the genome upon transformation. To investigate this possibility,

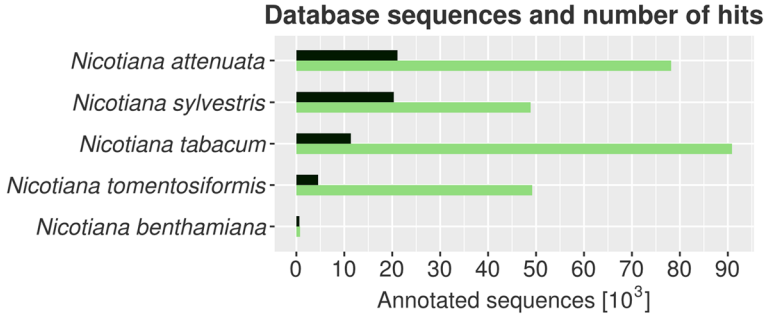


Fig. 1 Blast best hits of NibSet-1 proteins on the five most represented *Nicotiana* species in the database. Shown are the number of database sequences belonging to each species (green), and the number of hits that were used for functional annotation (black)

we generated Illumina paired-end genomic reads from the ΔXT/FT plant and from its wild-type parent, corresponding to 33-fold and 41-fold coverage, respectively, of the *N. benthamiana* genome (Additional file 2, code LF_DEX_3, LF_NIB_3). The transgenic constructs used in ΔXT/FT had a total length of 4.5 and 4.8 kbp, respectively, and were composed of the CamV35S promoter (2.8 kbp), the transgenic cassette (FucT-transgene, 1.1 kbp, or XylT-transgene, 0.8 kbp), and the 7TTR terminator region (0.9 kbp) [8]. We searched for the regions of the genome where the integration had taken place by identifying ΔXT/FT read pairs that had one mate mapping on the transgenic promoter or terminator sequence, respectively, and the other mate on the host genome represented by the Nb-1 draft assembly. For both transgenic constructs the whole sequence showed read coverage (Additional file 1: Figure S2), and we observed highly supported connections with Nb-1 scaffolds Niben101Scf03674 (62 pairs) and Niben101Scf03823 (32 pairs). We found promoter (P) and terminator (T) pairs clustering separately, defining the junction regions (Fig. 2). The clusters were composed of 34 P and 28 T pairs in Niben101Scf03674 and of 12 P and 20 T pairs in Niben101Scf03823. We note a difference between the two insertion sites in terms of number of bridging pairs. As outlined further below, the study of the insertion site in scaffold Niben101Scf03823 was problematic due to repetitive elements and assembly breakpoints. This likely reduced the ability of mapping reads to the region.

We performed a local alignment with the matching reads to localize the insertion position at base-pair precision by identifying chimeric reads that spanned the junctions between host genome and the transgenes. Supported by 10 P and 18 T chimeric reads we marked positions 27872 and 27901 as junction positions in Niben101Scf03674, and 11 P and 10 T chimeric reads supported positions 34601 and 41896 as junctions in Niben101Scf03823 (Fig. 2).

The location of mapped reads indicated that transgene integration in scaffold Niben101Scf03674 had led to a small deletion of 28 bases (Additional file 1: Figure S3).

In scaffold Niben101Scf03823 the context and the consequences of the insertion were less obvious (Fig. 2, panel "b", Fig. 3). The gap density in the insertion region, a high amount of annotated TEs, and a coverage drop in ΔXT/FT may support a scenario whereby the region was misassembled in the Nb-1 draft and altered by a rearrangement that took place during transgene insertion (see Additional file 1: text; Figure S4).

Molecular consequences of transgene insertions in ΔXT/FT

In the case of scaffold Niben101Scf03823, our data supported transgene insertion in a region consisting of non-coding, highly repetitive DNA, where no predicted gene was disrupted by the insertion. Therefore, this insertion site was considered as not critical regarding its functional impact. In contrast, the inferred insertion site in the region corresponding to scaffold Niben101Scf03674 was located within intron 4 of gene g76921, encoding for TFIID subunit 12-like isoform X1, a subunit of an important general transcription factor [31]. Analysing mRNA-seq data from ΔXT/FT (see below), the expression profile

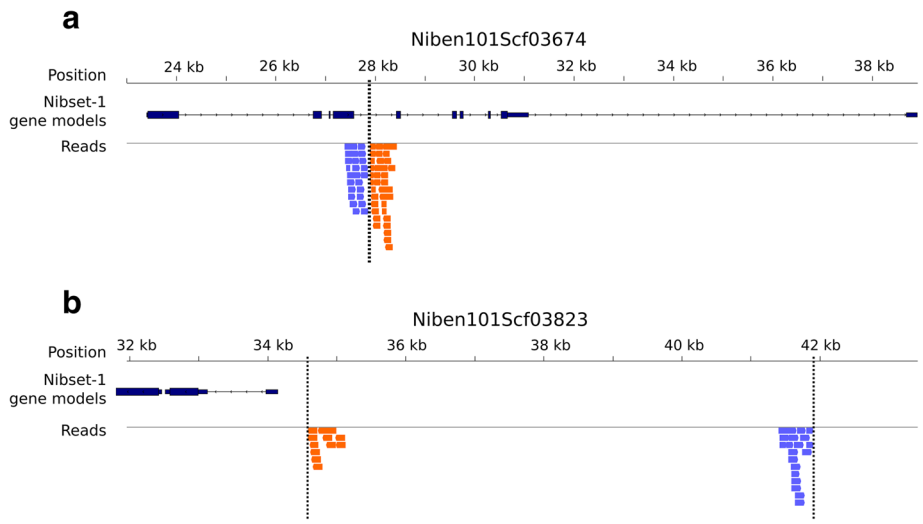


Fig. 2 Identification of read-pairs connecting *N. benthamiana* genomic sequence to transgene sequences in ΔXT/FT. Mates of a read pair which establish a connection with the promoter fragment of a transgene are shown in light blue, those indicating a connection with a transgene terminator sequence are shown in orange. NibSet-1 gene models identified within or in proximity of the insertion regions are shown in dark blue. Dashed lines represent junctions identified with chimeric reads. **a** Transgene integration site within scaffold Niben101Scf03674 at a position between exons 4 and 5 of NibSet-1 gene g76921. **b** Transgene insertion site at scaffold Niben101Scf03823

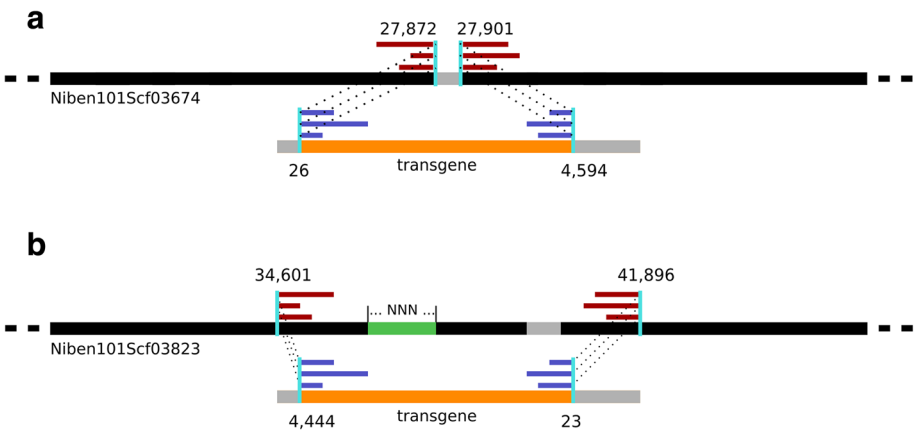


Fig. 3 Organisation of transgene integration sites within the genome of *N. benthamiana* ΔXT/FT, as explored using chimeric sequencing reads spanning integration boundaries. **a** Expected model for the chimeric reads mapping, as exemplified by the insertion at scaffold Niben101Scf03674. Shown are chimeric reads mapping on the host genome with one side (dark red) and on the transgene promoter/terminator site with the other side (blue). The dashed lines connect the two sides, while the cyan vertical lines outline the boundary between the host genome and the transgene. Regions without read coverage are shown in gray, while covered regions are shown in black (host) or orange (transgene). **b** Transgene insertion site on scaffold Niben101Scf03823. A stretch of undetermined bases (~ 500 bp) within the Nb-1 assembly is indicated in green. The numbers over the cyan vertical lines indicate the junction positions on the Niben101Scf03823 scaffold. The proposed rearrangement of the region is shown in Additional file 1: Figure S4

of this gene showed a much higher transcriptomic coverage in the exons downstream of the insertion site (exons 5–9) than in the exons further upstream (Fig. 4). This supported the idea that the transgene under the control of the CamV35S promoter had become fused to the exons of g76921 from exon 5 onwards in ΔXT/FT. Indeed, we found 11 transcriptomic read pairs that confirmed the occurrence of such a fusion transcript: these read pairs showed one mate mapping onto g76921 and the other mate mapping onto the FucT-transgene, unequivocally assigning its integration site to scaffold Niben101Scf03674. Therefore, we could infer that the XylT transgene insertion had occurred on scaffold Niben101Scf03823. However, no formal proof of this conclusion was possible due to highly repetitive sequences surrounding the integration site. Read pairs which linked the FucT transgene to g76921 mapped not only to exon 5 but also to exons 6 to 8, respectively, indicating that exons downstream of the insertion site kept their original splicing pattern. We concluded that the g76921 locus was disrupted in ΔXT/FT, and a fusion transcript composed of the FucT-transgene RNA attached to the normally spliced exons 5 to 9 of g76921 was present. Notably, we did not find read pairs linking exons 4 and exon 5 (i.e. no support for the presence of the wild type allele), indicating homozygosity, with both alleles of g76921 being disrupted. However, we considered a disruption of g76921 as not harmful to ΔXT/FT since there is another actively expressed gene copy annotated as TFIID subunit 12-like isoform X1 (g54961, 86% protein seq. Identity; Additional file 1: Figures S6, S7, S8). In principle, g54961 may be sufficient to buffer the loss of function of g76921; however, its TPM expression value in ΔXT/FT (12.6 ± 0.4) was comparable to the one observed in WT (13.8 ± 1.5) and the resulting log-2-fold change was negligible (– 0.029).

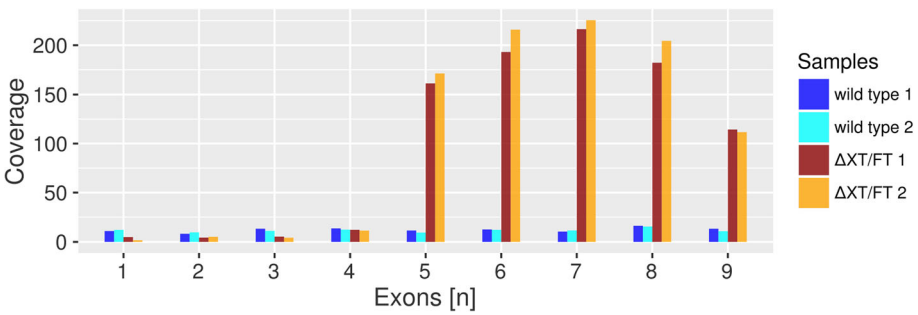


Fig. 4 Per-exon transcriptomic coverage of NibSet-1 gene g76921, as detected with transcriptomic reads. WT replicates are indicated in blue and cyan, and ΔXT/FT replicates in red and orange, respectively. Increased coverage of exons 5–9 in ΔXT/FT is the result of transgene integration which leads to the production of a fusion transcript under control of the strong CamV35S promoter

Analysis of the ΔXT/FT transcriptome

The perturbation of the ΔXT/FT genome upon transgene insertion might have unpredictable effects on the plant's transcriptome. We therefore generated leaf mRNA-seq data from ΔXT/FT and its wild type (WT) parent, both in duplicate. The paired-end reads were quality-trimmed and mapped against the Nb-1 draft genome assembly, using NibSet-1 gene models as guide for mapping. We extracted the raw counts for each gene in each replicate and condition; the counts were then normalized to the sequencing depth of the corresponding replicate. Genes with low mean coverage across replicates and samples (< 10) were removed. We assessed the potential presence of artifacts in the normalized counts through a principal component analysis (PCA). The PCA outlined no clear distinction between conditions and replicates (Additional file 1: Figure S9). Pearson's correlation scores calculated between the four samples were all ≥ 0.9 (Additional file 1: Table S8). We concluded that the transcriptome in WT and in ΔXT/FT are likely to be highly comparable. From the normalized counts of the retained genes we computed Fragments Per Kilobase of exon per Million fragments mapped (FPKM) and Transcripts Per Million (TPM) for each gene. We then computed log2-fold changes (LFC) between the two genotypes (Additional file 4). Considering the high correlation between the samples we made sure that even moderate variation in gene expression were considered; hence, we considered as differentially expressed every gene showing a LFC ≥ 0.5. The test returned a group of 21 differentially expressed genes (DEGs), all with LFC values substantially higher than the 0.5 threshold (≥ 1.40, Fig. 5). From this list we removed seven genes having a TPM value below the sample-specific TPM threshold (indicated in the Methods section) in both conditions.

We performed quantitative PCR in triplicate for the remaining 14 DEGs in order to confirm their differential expression. Unpaired *t* tests between ΔXT/FT and WT were performed to test the statistical robustness of each qPCR observation; we retained only those showing the same expression trend and a two-tailed *p*-value < 0.05. We confirmed one up-regulated gene (g76921) as well as three down-regulated genes (g10744, g25290, g29021) (Table 2, Fig. 6, Additional file 1: Figure S10). We note the presence of g76921 among the upregulated DEGs, which was disrupted by the insertion of the FucT-transgene (see above). Through interPro [32] we catalogued protein family, annotated domains, repeats, signature matches, and GO terms of the confirmed DEGs, none of them being directly involved in protein glycosylation. Notably, the four genes targeted by the transgenes (g31184, g80352, g43728, g40438) were not found among the five DEGs. This is most likely due to the efficiency of the knockdown system. We did, in fact,

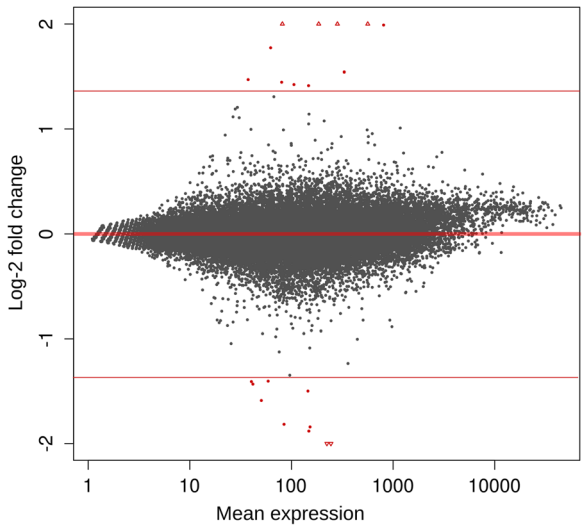


Fig. 5 Comparison of global gene expression in leaves of the *N. benthamiana* wild type WT and the transgenic line **Δ**XT/FT. Log2-fold change (LFC) is plotted against mean expression (counts) for each NibSet-1 gene. Red dots represent genes with a sufficient mean expression, LFC and adjusted *p*-value to be considered differentially expressed genes (DEGs). Triangles represent genes whose LFC value exceeded the plot margins [– 2,2]. Thin red lines enhance the separation between DEGs and the other genes. Eleven genes are found upregulated (log2-fold change > 1.40) and ten genes are found downregulated (log2-fold change < – 1.40) in **Δ**XT/FT compared to the *N. benthamiana* wild type

observe a generalized decrease in normalized read counts for the targeted genes in ΔXT/FT with respect to WT (Additional file 1: Table S4). We note that, while the transgenes were designed to act post-transcriptionally, potential homology of their promoter with that of other host genes could have triggered transcriptional gene silencing *in trans* [33–35], altering their transcription. As our results show that this was not the case, we conclude that ΔXT/FT has a transcriptional profile which is highly comparable to the wild type, with the exception of the transgene knockdown of FucT and XylT.

Genomic variants in ΔXT/FT

We screened the genome of *N. benthamiana* ΔXT/FT for differences (i.e. variants) that could have accumulated after the generation of ΔXT/FT, dated 2008, during at most 40–50 estimated generations by 2015, when the samples were taken and sequenced. The genotype Nb-1, an inbred *N. benthamiana* line that had been maintained in the laboratory of Gregory B. Martin since the mid-1990s [18] was used as a reference.

We re-sequenced the genomes of both ΔXT/FT and WT to approximately 33-fold and 41-fold respective genomic coverage on the Illumina sequencing platform (Additional file 2, codes LF_DEX_3 and LF_NIB_3) and used the reads to call variants relative to the Nb-1

Table 2 Differentially expressed genes (DEGs) between wild type *N. benthamiana* and the ΔXT/FT transgenic line based on a comparison of leaf mRNA-seq data and confirmation by quantitative PCR

Gene ID	Function	*E*-Value	Identity	TPM ΔXT/FT	TPM WT
Downregulated genes					
g10744	uncharacterized oxidoreductase At4g09670-like	0	96%	1.9 ± 0.6	18.5 ± 2.4
g25290	alpha-soluble nsf attachment	0	100%	4.3 ± 0.4	34.0 ± 1.1
g29021	PREDICTED: LOW QUALITY PROTEIN: primary amine oxidase-like	0	92%	0.1 ± 0.0	20.4 ± 1.3
Upregulated genes					
g76921	transcription initiation factor TFIID subunit 12-like isoform X1	0	85%	41.4 ± 2.8	5.1 ± 0.0

Gene IDs refer to NibSet-1. The protein sequences of the identified DEGs were mapped on the blast Eudicots database (taxid: 71240)

reference genome (see methods for details). To exclude consensus errors in the assembly, we mapped genomic reads from the Nb-1 genotype against the Nb-1 assembly and removed all varying positions from the analysis (Table 3, panel "a"). After this filtering step, 96,510 SNVs and 6,605 indels were detected between ΔXT/FT and Nb-1; 106,079 SNVs and 7,217 indels were detected between WT and Nb-1 (Table 3, panel "b"); in both cases a transition/transversion (Ti/Tv) ratio of 1.4 was observed. To obtain a list of ΔXT/FT specific variants, we removed 57,362 SNVs and 2,478 indels shared by both genotypes against the Nb-1 reference; In this way, 39,148 SNVs and 4,127 indels specific to ΔXT/FT were retained. Of these, 3,036 SNVs and 80 indels were found within coding regions (CDS) (Table 3, panel "b"). The Ti/Tv ratio within CDS was higher (1.8) than in the whole variant pool (1.4); this could be due to higher selective pressures against transversions in coding regions [36]. We annotated the impact of each variant with the program SnpEff [37] which returned 67 variants (23 SNVs, 44 indels) in different genes annotated as "high impact" variants (Additional file 5). We extracted GO terms for the proteins encoded by these genes, retrieving terms for 29 proteins (43.3%). However, with a false discovery rate (FDR) < 0.05, we found no statistically significant GO term enrichment.

Genetic relatedness of *N. benthamiana* research accessions

A recent study posits that today's laboratory strains of *N. benthamiana* are all derived from a single specimen collected in the central Australian desert [38, 39]. The two draft genome assemblies available [18, 20] diverge by one SNV every 2,900 base pairs, i.e. 345 SNV/Mbp [38]. To assess whether we could obtain comparable data based on coding regions, we selected seven *N. benthamiana* accessions from which public mRNA-seq data were available (Additional file 2), maintained at the following research institutions: China Agricultural University, Beijing, China; King Abdul Aziz University, Jeddah, Saudi Arabia; National Academy of Agricultural Sciences, Jeonju, South Korea; University of Sydney, Sydney, Australia; Swedish University of Agricultural Sciences, Uppsala, Sweden; University of Natural Resources and Life Sciences (BOKU),

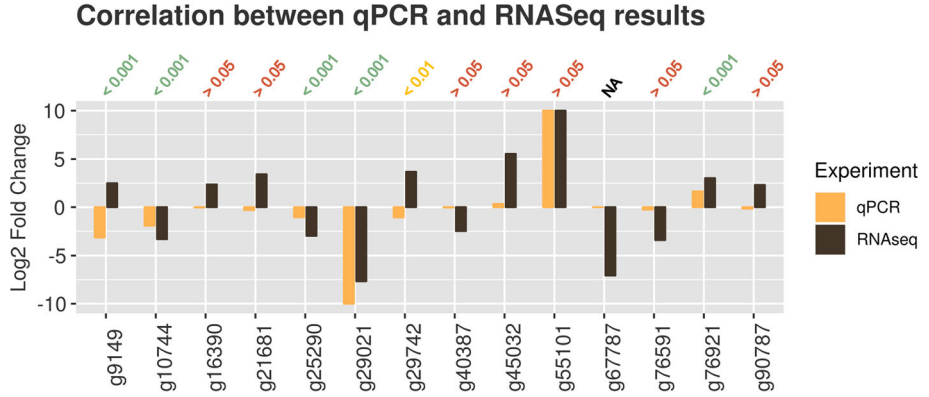


Fig. 6 Comparison of differential gene expression as assessed by mRNA-seq and quantitative PCR, respectively. Positive log2-fold change indicates upregulation in ΔXT/FT relative to wildtype, negative log2-fold change indicates downregulation in ΔXT/FT. Values above the table indicate two-tailed *p*-values computed in unpaired *t* tests between ΔXT/FT and WT qPCR results. Red: not significant, yellow: moderately significant, green: highly significant. Not available ("NA") indicates absence of qPCR signal. We kept for further analyses only genes where both qPCR and mRNA-seq indicated up-regulation or down-regulation, respectively. Gene numbers refer to NibSet-1 genes. The plot is limited to the range [− 10, + 10]

Table 3 Number of single-nucleotide variants (SNVs), number of insertion/deletion variants (indels) and transition/transversion (Ti/Tv) ratio for each comparison performed

Line	SNVs	Indels	Ti/Tv
a			
ΔXT/FT vs Nb-1	117,278	7,626	1.4
WT vs Nb-1	127,976	8,257	1.4
Nb-1 vs Nb-1	56,930	4,505	1.3
b			
ΔXT/FT vs Nb-1	96,510	6,605	1.4
WT vs Nb-1	106,079	7,217	1.4
Shared	57,362	2,478	1.4
ΔXT/FT unique	39,148	4,127	1.4
ΔXT/FT unique (CDS)	3,036	80	1.8

a) Raw number of variants before filtering out consensus errors, and b) after filtering out consensus errors, including subsets of variants relevant in the analysis. "Shared": variants shared between ΔXT/FT and WT relative to Nb-1. "ΔXT/FT unique": variants found only in ΔXT/FT relative to Nb-1. "ΔXT/FT unique (CDS)": variants found only in ΔXT/FT relative to Nb-1 restricted to coding regions

Vienna, Austria. From BOKU both the WT and ΔXT/FT accessions used in this study were included. We quality-trimmed reads from each accession, selected 14 million reads each and cropped them to a length of 48 nt. The number of reads extracted was chosen according to the maximum number available from each sample after quality filtering (smallest dataset: *N. benthamiana* accession from Jeonju, South Korea, 14 million reads). The cropping length was decided according to the longest common sequence length available after trimming (shortest reads: *N. benthamiana* accession from Uppsala, Sweden, 48 nt). As some of the datasets were single-end reads, the paired-end samples were processed using only the first read of each pair. The Nb-1 draft genome assembly was used as a reference for mapping.

For each obtained call set we computed the SNV/Mbp ratio dividing the number of SNVs by the positions (in Mbp) covered by the reads (min. Coverage 4x) limiting the computation to CDS regions only. All of the seven tested accessions showed similar rates, with an average of 67 SNV/Mbp (range: 64–75). The lowest recorded rate of SNV/Mbp belongs to the sample from Jeddah, Saudi Arabia, although we note that all of the values were in a very narrow range (Table 4). These values are compatible with the aforementioned divergence estimates by [38]: our estimates were obtained using coding regions, hence variation is expected to be lower than in whole-genome comparisons. The coding sequence-based divergence estimates are all very similar, supporting a scenario whereby the tested accessions display high genomic relatedness.

Table 4 Number of single-nucleotide variants (SNVs) obtained by mapping of mRNA-seq data from *N. benthamiana* and *N. sylvestris* against the Nb-1 reference genome sequence, considering only variants within coding exons

	Cov. Positions	SNVs	SNVs/Mbp
WT (AT)	8,630,008	556	64
ΔXT/FT (AT)	8,651,732	562	65
LAB (AU)	11,483,694	789	69
N. benthamiana (CN)	6,574,943	495	75
N. benthamiana (KR)	10,517,109	695	66
N. benthamiana (SA)	8,717,762	562	64
N. benthamiana (SE)	11,074,510	719	65
N. sylvestris	7,990,760	65,140	8,152

Covered positions: positions with a minimum coverage of 4x; SNVs: total number of variants detected in coding regions; SNVs/Mbp: number of variants per Megabase of coding sequence. Sample names are specified in the first column. Countries of origin are specified as follows: Australia (AU), Austria (AT), China (CN), Saudi Arabia (SA), South Korea (KR), Sweden (SE)

As a control, we used mRNA-seq reads from the presumable *N. benthamiana* paternal subgenome donor *N. sylvestris* [40] processed with the same pipeline; we obtained 8,152 SNV/Mbp distributed in 7,990,760 bp (Table 4). We also confirmed the validity of the variants within coding regions using contigs obtained by assembling ΔXT/FT genomic reads (see Additional file 1: Text). We observed a concordance of 84% between calls from mRNA-seq data (ΔXT/FT cDNA reads) and calls from contig mapping (124 mRNA-seq SNVs in agreement, 24 in disagreement).

As a means of comparison we analysed the variant density observed between *A. thaliana* accessions. For once, we called variants in annotated coding regions using mRNA-seq reads from six *A. thaliana* ecotype Col-0 derived lines in comparison to the TAIR10 reference genome assembly [41], using the same parameters as for *N. benthamiana*. Further, we used Col-0 mRNA-seq reads and mapped them against 13 different *Arabidopsis* genome assemblies of wild accessions generated in the 1001 genomes study [42]. Col-0 intra-accession diversity was very low (2 SNV/Mbp: range: 1–3 SNV/Mbp), while many more variants were observed in comparison to wild-derived accessions (1742 SNV/Mbp; range: 1447–2178 SNV/Mbp) (Table 5, panels "a" and "b").

Discussion

Providing a set of predicted genes along with a draft genome sequence increases greatly the molecular resources for further analyses of a species. Although the existing draft assembly of *N. benthamiana* was based only on short-read sequencing data we were able to predict a large proportion of full-length transcripts including start and stop codon. The gene set was established using comprehensive mRNA-seq data generated in this

Table 5 Number of single-nucleotide variants (SNVs) obtained by mapping of mRNA-seq data from *A. thaliana* against the TAIR10 reference genome sequence

	Cov. Positions	SNVs	SNVs/Mbp
a			
Col-0 (CN)	9,098,019	24	3
Col-0 (DE)	10,839,185	12	1
Col-0 (JP)	12,819,475	18	1
Col-0 (MX)	10,992,622	20	2
Col-0 (NL)	11,479,175	23	2
Col-0 (US)	12,320,980	21	2
b			
No-0 (DE)	13,205,980	22,006	1,666
Sf-2 (ES)	13,174,328	23,169	1,759
Can-0 (ES)	13,095,023	28,515	2,178
Edi-0 (GB)	13,198,944	22,051	1,671
Bur-0 (IE)	13,172,042	25,137	1,908
Ct-1 (IT)	13,207,544	23,498	1,779
Tsu-0 (JP)	13,205,663	21,836	1,654
Mt-0 (LY)	13,220,021	21,953	1,661
Kn-0 (LT)	13,185,117	23,141	1,755
Hi-0 (NL)	13,212,525	19,123	1,447
Ler-0 (PL)	13,216,378	21,857	1,654
Ws-0 (RU)	13,194,655	22,999	1,743

Only variants in coding exons were considered. Covered positions: positions with a minimum coverage of 4x; SNVs: total number of variants detected in coding regions; SNVs/Mbp: number of variants per Megabase of coding sequence. a) mRNA-seq data from *A. thaliana* ecotype Col-0 mapped against TAIR10. Provenance of each accession is indicated: China (CN), Taiwan (TW), Japan (JP), Mexico (MX), Netherlands (NL), United States of America (US). b) mRNA-seq data from Col-0 "NL" mapped on genome assemblies from thirteen different wild-derived *A. thaliana* accessions. Ecotype name and country of origin is indicated. Country codes: Germany (DE), Ireland (IE), Italy (IT), Japan (JP), Libya (LY), Lithuania (LT), Netherlands (NL), Norway (NO), Poland (PL), Russia (RU), Spain incl. Canary islands (ES), United Kingdom (GB)

study and validated by two independent approaches both demonstrating its high level of completeness. To avoid the inclusion of transposable elements we performed repeat masking and posterior filtering of predicted genes that overlapped with repeat annotations. In this way, we lost one of five described FucT genes in the final gene set although it had been predicted initially. Further genes may be filtered out similarly, however, the prediction procedure aimed for a minimized repeat content in the final gene set. The majority of our predicted *N. benthamiana* genes could be matched by functionally annotated genes from other species providing additional valuable information on the *N. benthamiana* gene set and validating the predictions once again. Complementing existing data of *N. benthamiana* we generated genomic sequencing data from two additional *N. benthamiana* accessions one of which was the engineered ΔXT/FT line. Two genomic regions of interest were analysed in detail, i.e. the insertion sites of transgenes for silencing of FucT and XylT genes involved in glycan addition to proteins. While the genomic locations of insertion and corresponding sequence scaffolds could be identified and assigned to each transgene we found a differing amount of genomic read data matching the two transgene insertion sites. This indicated a rather complex scenario for the insertion site of the XylT transgene including repetitive regions, genomic rearrangements, and a potential misassembly in Nb-1, all of which limited the mappability of sequencing reads. The FucT transgene insertion site was covered well by sequencing reads from the ΔXT/FT line revealing transgene insertion within a gene that most likely lost its function. Since another intact copy of a closely related homolog was detected in the genome no harmful effect is to be expected. Transcriptome analysis did not show remarkable differences between ΔXT/FT and the wild type demonstrating specific transgene activity. Further differences between the two lines were only minimal. When comparing several *N. benthamiana* lines used in research laboratories our data suggested that the *N. benthamiana* lab lines tested here were more closely related to each other than wild-derived *A. thaliana* accessions. At the same time, higher divergence existed between *N.benthamiana* lines in comparison to *A. thaliana* Col-0 derivatives. Even though *N. benthamiana* research strains have recently been reported to originate from one source [38, 39], to the best of our knowledge no effort has been made to preserve and maintain a genetically homogeneous strain as is the case for the *A. thaliana* Col-0 ecotype; this might result in the slightly higher variation among *N. benthamiana* accessions that we have observed. All in all, our data confirmed the hypothesis that all currently used *N. benthamiana* laboratory accessions derive from the strain collected at the Australian Granites site [38].

Conclusion

Over the years, the interest in *N. benthamiana* as an *in planta* protein expression platform has grown considerably, and much information has been accumulated. The gene set presented here, comprising 50,516 genes transcribed in 62,216 isoforms reflects this knowledge gain. However, our functional annotation results also show the lack of information still present: only 71% of the transcriptional isoforms could be functionally annotated. Further research will have to fill this information gap. Our study also showed the need for a genome and transcriptome analysis when using a transgenic plant: the identification of disrupted genes, their potentially altered expression, their copy number, and the zygosity of the insertion are important factors to detect any side-effects of the transgene insertion. The insertion sites of the two

transgenes in ΔXT/FT could be located, even though the position of only one insertion could be identified on the nucleotide level. In this study, we also addressed variation within the whole genome and within coding regions, respectively, as a mean to determine accession relatedness. We show that the variation within coding regions is compatible with a scenario whereby the LAB strain is at the root of all accessions used in *N. benthamiana* research [38].

Methods

Plant material and isolation of nucleic acids

Seeds of wild-type *Nicotiana benthamiana* plants originally described by Regner and co-workers [43] were provided by Herta Steinkellner (University of Natural Resources and Life Sciences, Vienna). *N. benthamiana* ΔXT/FT is regularly grown in the lab of co-author Richard Strasser who also developed the line [8]. Wild type and ΔXT/FT plants were grown on soil in a growth chamber at 22 °C with a 16-h-light/8-h-dark photoperiod. For extraction of nucleic acids, leaves from 5-week-old plants were immersed in liquid nitrogen and macerated with grinding balls in a mixer mill. Genomic DNA was isolated from 1.5 g leaves using a Nucleospin Plant II Maxi kit (Macherey-Nagel, Düren, Germany) according to the instructions of the manufacturer. RNA was isolated from 40 mg leaves using the SV Total RNA isolation kit (Promega, Madison, WI, USA).

Library preparation and Illumina sequencing

One microgram of genomic DNA was sheared in a S220 Focused-ultrasonicator (Covaris, Woburn, MA, USA) using covaris microtubes with a duty cycle of 10, intensity 5 and a cycle/burst of 200 for 35 s in order to achieve a peak fragment length of 700 bp. Genomic libraries were prepared using the NEBNext Ultra sample preparation kit (New England Biolabs, Ipswich, MA, USA) according to the recommendations of the manufacturer. Size selection of the libraries was performed on a 2% agarose gel with 1xTAE buffer. A gel slice containing the library fragments of interest was processed using the QIAgen gel extraction kit (Qiagen, Hilden, Germany) and further purified using QIAquick columns. Thereafter, the library was amplified using 7 cycles of PCR. Finally, the library quality was assayed on a DNA1000 chip using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). Library quantity was assessed on a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). From ΔXT/FT and from the corresponding wild type line, we obtained 414 million and 508 million raw read-pairs, respectively (Additional file 2, codes LF_DEX_3, LF_NIB_3). This translates into a genomic coverage of 33-fold (ΔXT/FT) and 41-fold (wild type), assuming a genome size of 3.1 Gbp.

mRNA-seq libraries were generated on a Tecan robotic workstation using the TruSeq stranded mRNA library prep kit (Illumina, San Diego, CA, USA) starting with 1 μg of total RNA. During RNA purification, genomic DNA was digested with RNase-free DNase I (Promega, Madison, WI, USA). Libraries were amplified using 15 PCR cycles. Library quality and quantity was assessed as above. Sequencing was performed in paired-end mode on the Illumina HiSeq 2500 with v4 sequencing chemistry using a 2 × 125 cycle protocol. We obtained between 28 and 38 million raw read-pairs per mRNA-seq library (Additional file 2, codes LF_DEX_1 and 2, LF_NIB_1 and 2).

Gene prediction

Raw reads (Additional file 2) were analyzed with FastQC [44]. Read trimming was conducted with Trimmomatic [45] (ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 AVGQUAL:30 MINLEN:36). The Nb-1 draft genome assembly [18] (v1.01, downloaded in January 2016) available at the SOL Genomics Network [19] was used as a reference for the mapping step. With RepeatModeler [46] (–engine ncbi) we generated a library of repetitive elements on this draft genome assembly. Only repeats belonging to the DNA elements, LTR, LINE, SINE, Helitron and Unclassified families were retained, in order to mask transposable elements which can interfere with gene prediction [47]. RepeatMasker [48] (–engine ncbi -gff -noisy -no_is -norna -nolow) was used to generate a masked version of the Nb-1 genome, together with an annotation in GFF format.

We mapped the transcriptomic reads (Additional file 2) to the Nb-1 draft assembly with BLAT [49] (–tileSize = 11 -minIdentity = 92 -stepSize = 11 -minMatch = 2 -maxGap = 2 -oneOff = 0) and with TopHat2 [50] (–-read-mismatches 2 --read-gap-length 2 --max-insertion-length 3 --max-deletion-length 3 --b2-sensitive --microexon-search). PCR duplicates were removed. The results were filtered with samtools [51] keeping only primary alignments (samtools view -F 0×0100). Expression hints from the mapping results of BLAT and TopHat2 were computed separately and combined, giving priority to TopHat2 results in case of conflicts. With the script RNA-seq-noise-reduction.pl [52] we increased the contrast between exon and intron regions. We further limited the hints coverage by applying a minimum coverage of 20 and a maximum coverage of 300 to each hint to reduce background noise. The combined mRNA-seq information was merged with the information on annotated repeats, yielding 72,940,895 hints for exonic positions (genome positions with mRNA-seq coverage), 583,572 hints for introns (full intron span defined by reads mapped in spliced mode) and 1,994,352 hints for repetitive sequences (from

RepeatMasker, see above). The unmasked Nb-1 draft genome assembly was split into 50 segments of similar size to parallelize the analysis. We provided repeat information in the hints file, instead of using the masked genome [52, 53]. Each segment was then submitted to the Augustus pipeline [26] (alternatives-from-evidence=true, allow-hinted-splicesites=atac, species=coyote_tobacco).

Gene set filtering and validation

The raw gene set generated by Augustus was filtered by removing gene structures with < 1% coverage by expression hints. We removed peptides of length < 10 amino acids from the protein set of sequences. We filtered out the genes that overlapped with annotated TEs by more than 10 nt in their coding regions. The consistency between mRNA-seq expression profiles and gene models was assessed for 200 randomly chosen genes with GBrowse2 [54] adding separate data tracks for expression evidence and for transposable elements. We assessed correlation between predicted exons and read coverage, between predicted introns and split-mapped reads, and the absence of annotated TEs in the coding regions. The Niben101_annotation gene set was downloaded from the SOL Genomics Network website (https://solgenomics.net/) [19], from the ftp repository corresponding to *N. benthamiana* (v101). The overlap between gene models was determined using bedtools intersect [55]. The concordance between annotated CDS regions was assessed with a custom Python script. The completeness of the gene set was verified with BUSCO [27] (–m OGS), using the BUSCO plant database (http://busco.ezlab.org/). To avoid biases in the duplicated BUSCOs counts we used only one sequence per gene, corresponding to its longest isoform. The BUSCO validation was run on both NibSet-1 and Niben101_annotation. *N. benthamiana* cDNA sequences were downloaded from GenBank [56]. The sequences were converted to protein sequences and mapped against the proteins of the newly generated gene set using BLAT [49] (–minIdentity=85). The PSL-formatted results were then filtered by sequence identity and alignment length.

Functional annotation

The validated gene set was functionally annotated using sequence homology. Four blast databases were built with the protein sequences belonging to the *Nicotiana* genus, to the *Solanaceae* family and to *A. thaliana*, downloaded from NCBI-Protein. The sequences were chosen by querying the NCBI-Protein database for the desired species, genus, family or group, including all the listed results. By generating taxonomically confined databases with significance for *N. benthamiana*'s phylogenetic history, we also reduced computational time. The blast databases were built with makeblastdb [57] (makeblastdb -dbtype prot -input_type fasta -parse_seqids). The preformatted non-redundant protein and non-redundant nucleotide databases were downloaded from the blast repository. We mapped the gene set encoded protein sequences against these databases with blastp [57] using default parameters and -evalue 0.001 -word_size 3 -outfmt 5 -max_target_seqs 1. The results were filtered keeping only alignments with an E-value ≤ 10e-10, an alignment length ≥ 70 amino acids, sequence identity ≥ 90% and an aligned sequence fraction ≤ 90% (Figs. 7 and 8). The aligned fraction of each sequence was computed with find-best-hit.py [58] which determines how much of the query sequence is covered by mutually compatible high scoring pairs (HSPs), i.e. by non-overlapping HSPs. We first mapped the protein sequences against the *Nicotiana* genus protein database. We then extracted the ones satisfying our criteria, and mapped the remainder against the *Solanaceae* protein database. This scheme was repeated, in order, with the *A. thaliana*, non-redundant protein and nucleotide databases. We did not consider as functionally annotated proteins with the descriptors "uncharacterized", "unknown", or "hypothetical" or proteins without a match.

Detection of transgene insertion sites

Raw genomic reads (Additional file 2) were inspected with FastQC [44]. Read trimming was conducted with Trimmomatic [45] (ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 AVGQUAL:30 MINLEN:36). We mapped ΔXT/FT paired-end genomic reads from a library with a peak insert size of 700 nt (Additional file 2, Barcode LF_DEX_3) against a combined reference that included the Nb-1 draft genome assembly and the two transgene insert sequences (XylT insert, 4,536 nt, FucT insert, 4,768 nt, both including the LB and RB sequences, Additional file 3) using HISAT2 [59] (hisat2 -I 500 -X 775 --no-spliced-alignment --score-min L,-0.6,-0.6 -k 2). We filtered the mapping results keeping primary alignments only (samtools view -F 0×0100). We then extracted read pairs with one mate mapping on an Nb-1 scaffold and the other mate mapping onto a transgene, labeling them as promoter (P) or terminator (T) pairs depending on which region of the transgene they were bridging; connections with < 10 bridging pairs were excluded from further analyses. Local mapping to detect chimeric reads was conducted with bwa [60] (bwa mem -m 5 -k 20 -c 10 -B 6 -O 5,5 -E 3,3 -U 0 -Y -T 20). We filtered the mapping results keeping primary alignments with supplementary alignments using samtools [51] (samtools view -f 2048 -F 0×0100). The junction positions were calculated from the leftmost mapping position, performing the CIGAR operations (BAM format, 6th field). Genomic read coverage per position was computed from the BAM file

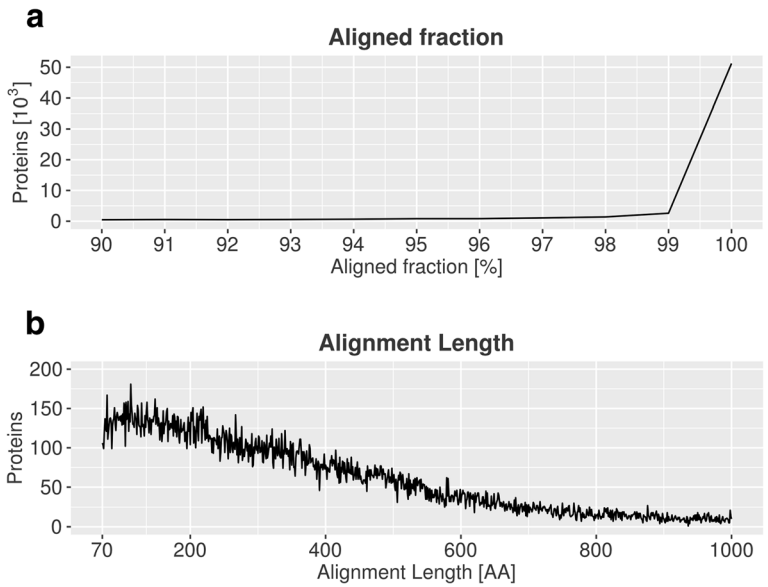


Fig. 7 Comparison of NibSet-1 protein sequences against protein sequence databases. **a** Fraction of the sequence of each protein in the NibSet-1 transcriptome aligned to its best match in one of the blast databases used in this study. **b** Length in amino acids of such alignments

used for the bridging pairs analysis, using samtools depth [51].

Gene disruption in ΔXT/FT

To search for fusion transcripts we concatenated the NibSet-1 transcriptome FASTA file with the two transgene cassette sequences (XylT, 840 nt; FucT, 1072 nt; both including sense, intron and antisense fragment). Trimmed transcriptomic reads from ΔXT/FT (Additional file 2) were used (trimming parameters see under "gene prediction"). We cropped the reads to a length of 36 nt to be able to map also most of the reads spanning the fusion junction; using end-to-end alignment those reads would not have aligned to the reference. We mapped the cropped reads with HISAT2 [59] (hisat2 --rdg 5,3 --rfg 5,3 -k 3 --no-spliced-alignment --no-softclip --ignore-quals

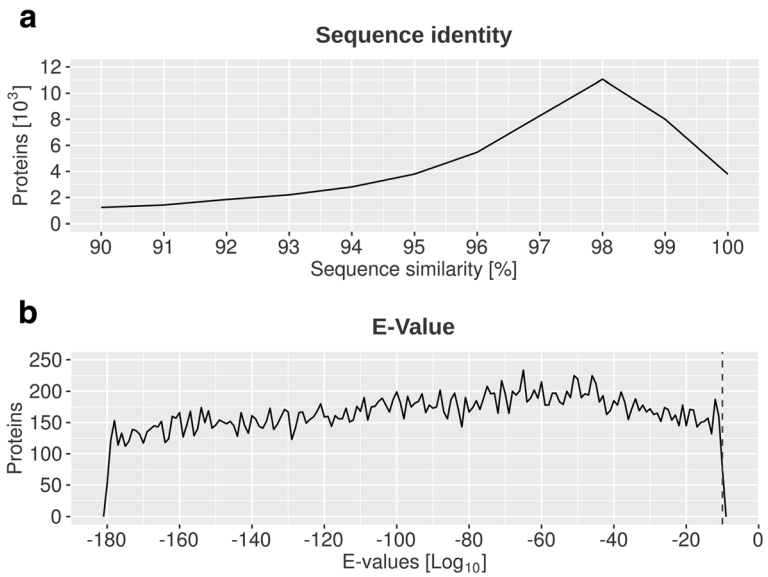


Fig. 8 Comparison of NibSet-1 protein sequences against protein sequence databases. **a** Sequence identity retrieved for each blast search. As the search was limited to results with at least 90% identity, the plot range is restricted from 90 to 100% identity. The y-axis shows the number of proteins (in thousands) matching at each level of identity. **b** Log10 of the e-values associated with each blast search. Results with e-values >10e-10 were discarded (dashed line)

--score-min L,-0.2,-0.3). We retained only primary alignments from the mapping results (samtools view -F 0×0100). We then extracted read pairs having one mate mapping on the transgene sense/antisense fragment ("insert mate"), and the other mate mapping on g76921 isoforms ("host mate"). The difference between the transgene cassette sequences allowed us to assign the FucT-transgene to this insertion site. Consequently, the XylT-transgene was assigned to the other. Transcriptomic coverage of g76921 was obtained with samtools depth [51], from the mapping scores of wild type and ΔXT/FT transcriptomic reads (Additional file 2).

ΔXT/FT expression profile

We mapped trimmed transcriptomic reads from ΔXT/FT and wild type with HISAT2 [59] (--mp 6,2 --rdg 5,3 --rfg 5,3 --score-min L,0.0,-0.2). We filtered the mapping results keeping primary alignments only (samtools view -F 0×0100) and computed read counts with HTSeq [61]. We expected the transcriptomic reads originating from transgenic molecules in ΔXT/FT to map on the regions they were designed to target. Hence, we filtered out read counts in the targeted regions of g31184, g40438, g43728 and g80352 (Additional file 1: Table S5) to avoid a bias in their log-2-fold changes (LFC) estimation caused by transgenic reads. We performed the principal component analysis (PCA) using the tools available within the DESeq2 package [62] and assessed Pearson's correlation coefficients using the R built-in cor function. We identified a list of differentially expressed genes (DEGs) with DESeq2 [62]. We kept only DEGs with an average mean coverage of at least 10 across replicates and conditions. We then tested for LFC ≥ 0.5 at $\alpha < 0.05$. For the resulting DEGs, we computed the TPM in each replicate and condition. We applied a sample-specific TPM threshold to consider a gene as expressed: we obtained the threshold via the conversion formula $TPM_i = (FPKM_i / sum_j(FPKM_j))*10^6$ [63] using $FPKM_i = 1$. Only genes with TPM equal or above threshold in at least one condition were kept. The thresholds used were 3.41, 3.43, 3.45 and 3.45 for samples LF_DEX_1, LF_DEX_2, LF_NIB_1 and LF_NIB_2 respectively. Function and GO terms for the identified DEGs were obtained by querying the online Eudicots database of Blast (taxid: 71240) [64] and interPro [32].

qPCR

Total RNA was reverse transcribed using the iScript cDNA Synthesis kit (Bio-Rad, Hercules, CA, USA). Real-time qPCR was performed in triplicate using the GoTaq qPCR master mix (Promega, Madison, WI, USA). Serine/threonine protein phosphatase 2A (PP2A) expression was used for normalization of qPCR data. Three independent biological replicates were used and mean values ± standard deviation are given, together with a two-tailed *p*-value representing the significance (Additional file 1: Figure S10). Primers used in this study are listed in Additional file 1: Table S6.

Genomic variants

Trimmed genomic sequencing reads (Additional file 2, codes LF_DEX_3, LF_NIB_3, trimming parameters see "Detection of the transgene insertion sites" methods section) were aligned to the Nb-1 draft genome assembly with Bowtie2 [65] (--sensitive --mp 6 --rdg 5,3 --rfg 5,3 --score-min L,-0.6,-0.6), setting a minimum and maximum insert size of 500 bp and 775 bp, respectively (–I 500 -X 775), which had been estimated by mapping a subset of 50,000 read pairs of each library (Additional file 1: Figure S11) against Nb-1. The used mapping parameters allowed a maximum of 12 mismatches, a maximum gap length of 23, or a combination of the two. The mapping returned a 21-fold coverage for ΔXT/FT and a 26-fold coverage for WT. The mapping results were then sorted by genomic coordinates keeping only the primary alignments (samtools view -F 0×0100). The raw call set was obtained with samtools mpileup [66] (call -f GQ,GP -v -m). Results were filtered with a combination of custom scripts. We required an average mapping quality and a calling quality of 20 (Phred score), a minimum coverage of 4, a maximum coverage of 30 for ΔXT/FT and of 38 for WT, a maximum fraction of reads with 0-mapping quality of 10% and a minimum number of reads per strand of 1. The filtered set of variants was compared with variants called with the same pipeline using sequencing reads isogenic to the plant used for the draft genome assembly (provided by A. Bombarely, Latham Hall, Virginia Tech, Blacksburg, VA, USA), to remove false calls due to consensus errors in the assembled genome. Isogenic sequencing reads were filtered with Trimmomatic using the following parameters: LEADING:25 TRAILING:25 SLIDINGWINDOW:4:20 AVGQUAL:35 MINLEN:40. Variants shared between ΔXT/FT and WT, and variants unique to either ΔXT/FT or WT were extracted with the bedtools "intersect" function [55].

The functional impact of variants annotated within coding regions of ΔXT/FT was assessed with SnpEff [37], identifying low, moderate and high impact variants as defined in the program documentation (http://snpeff.sourceforge.net/SnpEff_manual.html#eff). We performed a GO term analysis for the genes containing a variant with high impact. This analysis was conducted with InterproScan [67].

Transcriptomic variants

Quality-filtered reads from *N. benthamiana* samples ΔXT/FT and WT, *N. benthamiana* samples from

research institutions other than BOKU (SRR651957, SRR2976595, ERR219219, SRR1043177, SRR2085476), *N. sylvestris* (ERR274390) and *A. thaliana* (SRR6236990, SRR5195552, SRR3223423, SRR3928353, SRR5040365, DRR070513) were cropped to a length of 48 nt. *N. benthamiana* and *N. sylvestris* reads were downsampled to 14 million reads, while *A. thaliana* reads were downsampled to 8.5 million reads. Reads were mapped against the Nb-1 draft genome assembly [18] with HISAT2 [59] (--trim5 5 --no-softclip --mp 6,6 –rdg 5,3 –rfg 5,3 --score-min L,2.4,-0.3). Only primary alignments (samtools view -F 0×0100) mapping within CDS regions (i.e. excluding UTRs) were retained, if they had at least one mismatch difference between primary and secondary alignment; PCR duplicates were removed with Picard (http://Broadinstitute.Github.Io/Picard). Coverage was extracted with samtools depth [51]. Candidate variants were obtained through samtools mpileup [66] (–t DP, AD,ADF,ADR,SP,DP4) and bcftools call [68] (–f GQ,GP -v -m). We excluded: positions within 10 nt from an indel; indels within 100 nt from each other; clusters of 3 SNVs within 10 nt (all likely alignment artifacts). We requested a minimum base quality of 20, a minimum average mapping quality of 20, a minimum coverage of 4x, a minimum fraction of 0.1 (10%) reads with 0-mapping quality (MQ0F), a minimum fraction of 0.9 (90%) reads showing the alternative allele at each variant position. The thirteen different assemblies of *A. thaliana* were downloaded from the 1001genomes website [42]. For each we determined the coding regions by mapping the TAIR10 [41] *A. thaliana* transcript sequences against the assemblies with GMAP [69] (–f gff3_gene --min-identity 0.95); CDS lines from the resulting GFF3 file were piped to bedtools merge [55] to generate a non-redundant representation of coding positions. Reads from the "Netherlands" sample (lab-grown ecotype Col-0) were mapped against each of the assemblies, and variants were called using the same programs and criteria as used for the six Col-0 accessions.

Additional files

Additional file 1: **Table S1.** Transposable elements within the *N. benthamiana* reference genome. **Table S2.** BUSCO analysis to assess gene set completeness. **Table S3.** Number of sequences, database total length of each constructed database. **Table S4.** Normalized counts for target genes of the FucT and XylT transgenes. **Table S5.** Regions of FucT1, FucT-pseudogene, XylT1, XylT2 targeted by transgenes. **Table S6.** Primer sequences for qPCR. **Table S7.** Potential off-target effects of FucT-transgene and XylT-transgene. **Table S8.** Pearson's correlation between normalized counts of the four mRNA-seq samples. **Figure S1.** Gene models obtained by mapping sequences of FucT and XylT genes onto the Nb-1 draft genome assembly. **Figure S2.** Genomic coverage of transgenes within the ΔXT/FT genome. **Figure S3.** Genomic coverage in ΔXT/FT and wild type on scaffold Niben101Scf03674 and Niben101Scf03823. **Figure S4.** Re-assembly of region of insertion of XylT transgene. **Figure S5.** Alignment between scaffolds containing genes g76921 and g54961. **Figure S6.** Protein sequence alignment between genes g76921 and g54961. **Figure S7.** Multiple sequence alignment of g76921 and g54961. **Figure S8.** Folding of *N. benthamiana* proteins encoded by g76921 and g54961. **Figure S9.** Principal component analysis (PCA) on normalized read counts. **Figure S10.** ΔΔCT values and TPM for differentially expressed genes. **Figure S11.** Insert size estimation of ΔXT/FT, WT genomic sequencing libraries. (PDF 8074 kb)

Additional file 2: High-throughput sequencing data. (ODS 25 kb)

Additional file 3: Sequences of the constructs used for the generation of ΔXT/FT. (PDF 168 kb)

Additional file 4: Differential gene expression analysis results. (TXT 4526 kb)

Additional file 5: High impact variants. (TXT 441 kb)

Acknowledgements
We thank Ulrike Vavra and Christiane Veit for performing RNA and genomic DNA isolation as well as qPCR experiments. We are also grateful to Aureliano Bombarely for providing Nb-1 sequencing reads. Sequencing was performed at the Genomics Unit of the Centre for Genomic Regulation (CRG) in Barcelona, Spain.

Authors' contributions
HH, JCD and LM conceived and supervised the study. MS performed computational analyses. RS provided important reagents and supervised experimental work. MS, HH and JCD wrote the manuscript. All authors read and approved the final version of the manuscript.

Funding
This work was funded by the Austrian Science Fund FWF (Doctoral program BioToP, Project W1224).

Availability of data and materials
N. benthamiana genomic and transcriptomic data from BOKU wild type and ΔXT/FT lines are available under SRA Bioproject PRJNA481441, accession numbers SRR7540369, SRR7540370, SRR7540371, SRR7540372, SRR7540367, SRR7540368. Gene models, predicted protein sequences and gff files are available at http://bioinformatics.boku.ac.at/NicBenth/Download/. The sequence of the plasmid used in the generation of ΔXT/FT is provided as Additional file 3. *N. benthamiana* seeds can be obtained from one of the authors (RS). The transgenic line ΔXT/FT is available for academic research upon signature of a Material Transfer Agreement.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biotechnology, University of Natural Resources and Life Sciences (BOKU), Muthgasse 18, 1190 Vienna, Austria. [2]Department of Applied Genetics and Cell Biology, University of Natural Resources and Life Sciences (BOKU), Muthgasse 18, 1190 Vienna, Austria.

Received: 7 March 2019 Accepted: 8 July 2019
Published online: 19 July 2019

References
1. Fischer R, Emans N. Molecular farming of pharmaceutical proteins. Transgenic Res. 2000;9(4–5):279–99 discussion 277.
2. Schillberg S, Emans N, Fischer R. Antibody molecular farming in plants and plant cells. Phytochem Rev. 2002;1(1):45–54.
3. Goodin MM, Zaitlin D, Naidu RA, Lommel SA. *Nicotiana benthamiana*: its history and future as a model for plant–pathogen interactions. Mol Plant-Microbe Interact. 2008 Apr;2008(1):28–39.

4. van Herpen TWJM, Cankar K, Nogueira M, Bosch D, Bouwmeester HJ, Beekwilder J. *Nicotiana benthamiana* as a production platform for artemisinin precursors. PLoS One. 2010 Dec 3;5(12):e14222.
5. Castilho A, Bohorova N, Grass J, Bohorov O, Zeitlin L, Whaley K, et al. Rapid high yield production of different Glycoforms of Ebola virus monoclonal antibody. PLoS One. 2011;6(10):e26040.
6. Li J, Stoddard TJ, Demorest ZL, Lavoie P-O, Luo S, Clasen BM, et al. Multiplexed, targeted gene editing in *Nicotiana benthamiana* for glyco-engineering and monoclonal antibody production. Plant Biotechnol J. 2016; 14(2):533–42.
7. Kim M-Y, Van Dolleweerd C, Copland A, Paul MJ, Hofmann S, Webster GR, et al. Molecular engineering and plant expression of an immunoglobulin heavy chain scaffold for delivery of a dengue vaccine candidate. Plant Biotechnol J. 2017;15(12):1590–601.
8. Strasser R, Stadlmann J, Schähs M, Stiegler G, Quendler H, Mach L, et al. Generation of glyco-engineered *Nicotiana benthamiana* for the production of monoclonal antibodies with a homogeneous human-like N-glycan structure. Plant Biotechnol J. 2008;6(4):392–402.
9. Dent M, Hurtado J, Paul AM, Sun H, Lai H, Yang M, et al. Plant-produced anti-dengue virus monoclonal antibodies exhibit reduced antibody-dependent enhancement of infection activity. J Gen Virol. 2016;97(12):3280–90.
10. Montero-Morales L, Maresch D, Castilho A, Turupcu A, Ilieva KM, Crescioli S, et al. Recombinant plant-derived human IgE glycoproteomics. J Proteome. 2017;161:81–7.
11. Shental-Bechor D, Levy Y. Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. Proc Natl Acad Sci. 2008;105(24):8256–61.
12. Skropeta D. The effect of individual N-glycans on enzyme activity. Bioorg Med Chem. 2009;17(7):2645–53.
13. Lisowska E. The role of glycosylation in protein antigenic properties. Cell Mol Life Sci. 2002;59(3):445–55.
14. Dowling W, Thompson E, Badger C, Mellquist JL, Garrison AR, Smith JM, et al. Influences of glycosylation on antigenicity, immunogenicity, and protective efficacy of Ebola virus GP DNA vaccines. J Virol. 2007;81(4):1821–37.
15. Bosch D, Castilho A, Loos A, Schots A, Steinkellner H. N-glycosylation of plant-produced recombinant proteins. Curr Pharm Des. 2013;19(31):5503–12.
16. Rup B, Alon S, Amit-Cohen B-C, Brill Almon E, Chertkoff R, Tekoah Y, et al. Immunogenicity of glycans on biotherapeutic drugs produced in plant expression systems-the taliglucerase alfa story. PLoS One. 2017;12(10):e0186211.
17. Castilho A, Gruber C, Thader A, Oostenbrink C, Pechlaner M, Steinkellner H, et al. rocessing of complex N-glycans in IgG Fc-region is affected by core fucosylation. mAbs. 2015;7(5):863–70.
18. Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Mol Plant-Microbe Interact. 2012;25(12):1523–30.
19. Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, Strickler SR, et al. The sol genomics network (SGN)--from genotype to phenotype to breeding. Nucleic Acids Res. 2015;43(D1):D1036–41.
20. Naim F, Nakasugi K, Crowhurst RN, Hilario E, Zwart AB, Hellens RP, et al. Advanced engineering of lipid metabolism in *Nicotiana benthamiana* using a draft genome and the V2 viral silencing-suppressor protein. PLoS One. 2012;7(12):e52717.
21. Nakasugi K, Crowhurst RN, Bally J, Wood CC, Hellens RP, Waterhouse PM. De novo transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*. PLoS One. 2013;8(3):e59534.
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. 2014;9(3):e91776.
23. Kourelis J, Kaschani F, Grosse-Holz FM, Homma F, Kaiser M, van der Hoorn RAL. Re-annotated gene models for enhanced proteomics and reverse genetics. bioRxiv. 2018 [cited 2019 Feb 15]; Available from: http://biorxiv.org/lookup/doi/10.1101/373506
24. Wessler S. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev. 1995;5(6):814–21.
25. Bennetzen JL. Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. Plant Cell. 2000;12(7):1021–30.
26. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Web Server issue):W435–9.
27. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
28. NCBI Resource Coordinators. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2017;45(D1):D12–7.
29. Jansing J, Sack M, Augustine SM, Fischer R, Bortesi L. CRISPR/Cas9-mediated knockout of six glycosyltransferase genes in Nicotiana benthamiana for the production of recombinant proteins lacking β-1,2-xylose and core α-1,3-fucose. Plant Biotechnol J. 2019;17(2):350–61.
30. Zeng F-S, Zhan Y-G, Zhao H-C, Xin Y, Qi F-H, Yang C-P. Molecular characterization of T-DNA integration sites in transgenic birch. Trees. 2010; 24(4):753–62.
31. Burley SK, Roeder RG. Biochemistry and structural biology of transcription factor IID (TFIID). Annu Rev Biochem. 1996;65(1):769–99.
32. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009; 37(Database):D211–5.
33. Fojtová M, Bleys A, Bedřichová J, Van Houdt H, Křížová K, Depicker A, et al. The trans-silencing capacity of invertedly repeated transgenes depends on their epigenetic state in tobacco. Nucleic Acids Res. 2006;34(8):2280–93.
34. Park Y-D, Papp I, Moscone EA, Iglesias VA, Vaucheret H, Matzke AJM, et al. Gene silencing mediated by promoter homology occurs at the level of transcription and results in meiotically heritable alterations in methylation and gene activity. Plant J. 1996;9(2):183–94.
35. Thierry D, Vaucheret H. Sequence homology requirements for transcriptional silencing of 35S transgenes and post-transcriptional silencing of nitrite reductase (trans)genes by the tobacco 271 locus. Plant Mol Biol. 1996;32(6):1075–83.
36. Lyons DM, Lauring AS. Evidence for the selective basis of transition-to-Transversion substitution Bias in two RNA viruses. Mol Biol Evol. 2017;34(12): 3205–15.
37. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w ^{1118} ; iso-2; iso-3. Fly. 2012;6(2):80–92.
38. Bally J, Jung H, Mortimer C, Naim F, Philips JG, Hellens R, et al. The rise and rise of *Nicotiana benthamiana*: a Plant for all Reasons. Annu Rev Phytopathol. 2018;56(1):405–26.
39. Bally J, Nakasugi K, Jia F, Jung H, Ho SYW, Wong M, et al. The extremophile Nicotiana benthamiana has traded viral defence for early vigour. Nat Plants 2015;1:15165.
40. Kelly LJ, Leitch AR, Clarkson JJ, Knapp S, Chase MW. Reconstructing the complex evolutionary origin of wild allopolyploid tobaccos (*Nicotiana* section *Suaveolentes*). Evolution. 2012;67(1):80–94.
41. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40(D1):D1202–10.
42. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 2016;166(2):481–91.
43. Regner F, da Câmara MA, da Câmara Machado ML, Steinkellner H, Mattanovich D, Hanzer V, et al. Coat protein mediated resistance to plum pox virus in Nicotiana clevelandii and N. benthamiana. Plant Cell Rep. 1992; 11(1):30–3.
44. Andrews S. Fastqc: a quality control tool for high throughput sequence data [internet]. 2011. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
45. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
46. Smit A, Hubley R. RepeatModeler [internet]. 2008. Available from: http://www.repeatmasker.org
47. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(5):329–42.
48. Smit A, Hubley R, Green P. RepeatMasker [Internet]. 2013. Available from: http://www.repeatmasker.org
49. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
50. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.
51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
52. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015 Sep 2;16:184.

53. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2013;505(7484):546–9.
54. Stein LD. Using GBrowse 2.0 to visualize and share next-generation sequence data. Brief Bioinformatics. 2013;14(2):162–71.
55. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
56. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2017;45(D1):D37–42.
57. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
58. Wintersinger J. find-best-hit.py [Internet]. 2014. Available from: https://gist.github.com/jwintersinger
59. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
60. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010;26(5):589–95.
61. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
62. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
63. Pachter L. Models for transcript quantification from RNA-Seq. ArXiv e-prints. 2011;1104:3889 Available from: https://arxiv.org/pdf/1104.3889.pdf.
64. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
65. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
66. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
67. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
68. Danecek P, McCarthy SA, HipSci Consortium, Durbin R. A Method for Checking Genomic Integrity in Cultured Cell Lines from SNP Genotyping Data. PLoS ONE. 2016;11(5):e0155014.
69. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.

Publisher's Note

Supplementary Material

1 index

Supplementary Text:

2 *ΔXT/FT XylT-transgene insertion site assembly*
2 *Possible complementation of TFIID subunit 12-like isoform X1 gene disruption in ΔXT/FT*
3 *ΔXT/FT transgene targets and off-targets*
4 *Transgene-targeted regions*
4 *Genomic variant calling with assembled contigs*

5 Supplementary methods

7 Supplementary Tables S1-S8

11 Supplementary Figures S1-S11

20 Supplementary Table Legends

21 Supplementary Figure Legends

23 Supplementary References

ΔXT/FT XylT-transgene insertion site assembly

To get more information on the sequences flanking the transgene ends, we re-assembled the insertion region using ΔXT/FT genomic Illumina sequencing reads. We assembled the transgene promoter with 570 bp of flanking genomic sequence, and the terminator together with 1451 bp of flanking genomic sequence. The flanking regions corresponded to the sequences where bridging pairs mapped with their host mate (Figure 2, panel “b”). We mapped genomic paired-end reads on the reassembled version of the region, to detect candidate links to other contigs. We formed a connection upon detection of at least 10 bridging pairs. Each connected contig was then mapped against the Nb-1 draft genome assembly to identify its position with respect to the identified junctions. Connections were evaluated manually. The proposed structure is the one which required the least number of assumptions (Supplementary Figure S4).

Possible complementation of TFIID subunit 12-like isoform X1 gene disruption in ΔXT/FT

Considering that *N. benthamiana* is an ancient tetraploid, we assessed whether other copies of a related TFIID subunit 12-like isoform X1 gene could compensate for a loss of function at the g76921 locus. We aligned the protein sequence of g76921 against the NibSet-1 protein sequences, and detected a match at 90% minimum identity with g54961, also annotated as TFIID subunit 12-like isoform X1. To confirm that the two genes are located on different chromosomes and do not represent the same gene separately assembled twice, we aligned against each other the two scaffolds that contain the genes g76921 and g54961, respectively (Supplementary Figure S5). The alignment showed that these scaffolds do not represent separately assembled haplotypes.
We then checked whether g54961 is an active gene. We observed high expression (FPKM=9.0, TPM=13.9) in leaf tissue and did not detect differential expression (LFC = -0.02) between ΔXT/FT and WT; we therefore concluded that g54961 is expressed in ΔXT/FT as it is in WT. We then performed a pairwise global alignment between the g76921.t1 and g54961.t1 protein sequences (Supplementary Figure S6). The functional domain that is specific to the TFIID 20 kDa protein family (PF03847) of Pfam (70) was found with 57-60% sequence identity in the two sequences (positions 394-461 in g76921.t1, positions 408-475 in g54961.t1); the domain matched with bitscore 102.0 and E-Value 1.7e-29 on g76921.t1, with bitscore 106.7 and E-Value 5.8e-31 on g54961.t1.
We then assessed the level of conservation between TFIID subunit 12-like isoform X1 proteins across species. We downloaded the protein sequences annotated as such in the NCBI-protein database and performed a multiple sequence alignment including the two *N. benthamiana* proteins g76921.t1 and

g54961.t1. The results indicate that these protein sequences are not well conserved, in contrast to the C-terminal TFIID-specific well-conserved functional domain (Supplementary Figure S7). We finally performed a secondary structure prediction (71) that showed a strong alpha-helix propensity in this region. Using Swiss-MODEL (72,73) we observed the predicted fold for both g76921.t1 and g54961.t1, demonstrating that the folding of both proteins is essentially equal (Supplementary Figure S8). We concluded that the disruption of g76921 likely has no impact on the plant and that g54961 could potentially buffer the loss of function of the disrupted gene.

ΔXT/FT transgene targets and off-targets

The two transgenes were designed to target the core α1,3-fucosyltransferases (FucT) and β1,2-xylosyltransferases (XylT) of *N. benthamiana*. For the FucT gene family, five copies have been reported in *N. benthamiana* (29), with two of them characterized in a previous study (8). The two copies characterized in Strasser et al. (2008) are the ones targeted by the FucT-transgene that was inserted in the ΔXT/FT genome; they are named FucT1 and FucT-pseudogene (FucT-p), and the latter is not encoding a functional protein. They are highly similar in terms of sequence and a single transgene was used to target both of them. The other transgene was used to target the two documented XylT genes (XylT1 and XylT2). The mRNA sequences of these genes are publicly available in Genbank (EF562630.1, EF562631.1, EF562628.1, EF562629.1), and were obtained by sequencing cDNA from the wild type genotype used for the generation of ΔXT/FT. Based on sequence identity, we identified the four corresponding transcripts within gene set NibSet-1 (FucT1: g31184.t1, FucT-p: g80352.t1, XylT: g43728.t1, XylT: g40438.t1). NibSet-1 and NCBI gene sequences matched with ≥ 99% sequence identity. We concluded that these were the four targeted transcripts in ΔXT/FT. Based on this and other observations, we replaced the NibSet-1 sequences of these four genes by their manually curated Genbank counterparts.

We then assessed whether the transgenes could have an off-target effect on other transcripts. The hairpin RNAs produced by the transgenes are processed into short RNA molecules with a length of 19-30 nt (74). The processed RNA molecules interfere with target transcripts either as siRNAs, mediating cleavage of the target (75), or as miRNAs through mechanisms that prevent protein translation (76). Optimal criteria for gene silencing have been assessed multiple times, leading to different conclusions (77–80). In any case, the silencing efficiency correlates with the number of short RNAs that have 100% sequence identity with their target. We therefore generated all possible k-mers of length 19 nt, 21 nt, 25 nt and 30 nt, respectively, from the sequences of the two transgene fragments in sense orientation (Supplementary Table S7) and mapped each k-mer set against the

NibSet-1 transcripts with BLAT, requiring a sequence identity of 100% and allowing multiple hits for each k-mer. With 19-mers we observed two putative off-target transcripts (g3481.t1 and g36277.t1); both were annotated as “PREDICTED: glycoprotein 3-alpha-L-fucosyltransferase A-like” in our functional annotation, and correspond to the FucT3 and FucT5 genes reported in Jansing et al. (2018), respectively. These two were not found among the results obtained with the longest k-mers (30-nt); also, the difference in the number of matching k-mers between the targeted genes and the two putative off-targets was large (Supplementary Table S7). This is in line with the findings in Jansing et al. (2018), as in such study, another transgene had to be generated in order to target the additionally characterized FucT copies. We concluded that there likely are no off-target effects with the two transgenes used.

Transgene-targeted regions

We determined regions of the transcripts that are targeted by the transgenes. We mapped the sense fragment of the transgenes (FucT-transgene=426 nt, XylT-transgene=314 nt, both excluding the leading 'TCTAGA' XbaI restriction site, see Supplementary File 2) against the four NCBI transcript sequences and against their four NibSet-1 counterparts (Supplementary Table S5) using Blast. We observed high identity matches with FucT1, FucT-p and XylT1 for both the NCBI and the NibSet-1 version, but no match with XylT2 in the NibSet-1 (i.e. Augustus) version (g40438.t1, Supplementary Table S5). We therefore assessed the differences between the NibSet-1 and the NCBI sequences. We first obtained gene models in GFF3 format from the NCBI sequences by mapping them on the Nb-1 draft genome assembly with GMAP. We then visualized them together with the NibSet-1 models in a genome browser and we assessed the differences. The models showed overall compatible exon-intron junctions between NibSet-1 and NCBI; the differences resided mostly in the beginning or the end of each model (Supplementary Figure S1). In g40438, the transgene-targeted region was found in a part of the 3' end that is not present in the NibSet-1 model (Supplementary Figure S1, panel “d”). For the other three targeted genes, the regions targeted by the transgene were included in both models.

Genomic variant calling with assembled contigs

The ΔXT/FT genomic paired-end reads used for variant calling were assembled using SOAP-denovo2, with k-mer size 80. We generated 697,264 unscaffolded contigs of at least 1,000 bp length, with a contig N50 of 2,904 bp (total length of assembly with contigs ≥ 1000 bp: 1.92 Gbp). We mapped contigs of at least 1000 bp against the Nb-1 draft genome assembly using nucmer (81) requesting a

minimum alignment length of 1000 nt and a minimum alignment seed of 30 nt. Given the high repeat content and the allopolyploidy of the *N. benthamiana* genome, we allowed nucmer to extend only seeds that were unique both in the genome and in the mapped contig (--mum option). 691,656 contigs (99.2%) found a mapping location of which 604,358 were mapped uniquely. We extracted the mismatching positions from the filtered mapping scores and generated a VCF-formatted file containing candidate SNVs. We note that a higher number of variants is usually recovered with this method, as a fully sensitive alignment of a long sequence will detect many SNVs that are missed by short reads. When asking a contig to map with a minimum identity of 75% and at least 10% of its sequence mapping uniquely, we obtained 2743 SNVs/Mbp over the whole genome and 601 SNVs/Mbp over coding sequences (CDS). In both cases we divided the SNVs by the positions covered by the contigs.

Supplementary Methods

Transgene insertion site assembly

To assemble the insertion region in scaffold Niben101Scf03823, we extracted matching Illumina reads from ΔXT/FT with samtools view. With bedtools bamtofastq (69) we generated paired FASTQ files based on the mapping scores recorded in the BAM files. Unpaired reads were saved in a separate FASTQ file. The generated FASTQ files were used as input for SOAP-denovo2 (82) (all -w -L 450 -K 57 -d 1 -D 1 -F -w -G 70 -L 200 -c 5 -C 35 -b 1.75 -M 3). We mapped the known promoter/terminator sequences against the generated assembly with BLAT (-minIdentity=95), identifying the scaffolds corresponding to the promoter or to the terminator sequence, respectively. We mapped these scaffolds against the Nb-1 assembly with BLAT (-minIdentity=95) to identify the location of the co-assembled flanking regions. To extend the co-assembled flanking regions we mapped the genomic paired-end reads of ΔXT/FT against the generated assembly using the same pipeline used for the bridging pairs analysis.

Analysis of protein structure using web resources

To find genes similar to g76921 within the *N. benthamiana* genome, we aligned the g76921 protein sequence against the NibSet-1 predicted protein sequences with BLAT (-prot -minIdentity=90). Pairwise alignment between protein sequences was done on the EMBOSS Needle web server (83). The multiple sequence alignment was performed on the MUSCLE web server (84). The secondary structure prediction was done according to the Chou & Fasman algorithm using the APSSP web

server, unpublished but referenced in CAFASP3 (85). The protein fold prediction was obtained on the Swiss-MODEL web server. Genomic sequence comparison between scaffolds were performed using nucmer (-b 200 -c 65 --delta -g 1000 -l 10000).

ΔXT/FT transgene targets and possible off-target effects

We mapped the FucT1, FucT-pseudogene, XylT1 and XylT2 mRNA sequences deposited in Genbank (EF562630.1, EF562631.1, EF562628.1, EF562629.1) against NibSet-1 with BLAT (-minIdentity=80). We mapped the sense fragments of the transgenes (FucT-transgene, 426 nt; XylT-Transgene, 314 nt; see Supplementary File 2) against the mRNA sequences from Genbank (see above) and against the corresponding NibSet-1 genes using Blast (match/mismatch 1/-2, word size 28). We generated gene models for the targeted transcripts with GMAP (-f gff3_gene --min-identity 0.95). We generated all the possible k-mers of sizes 19, 21, 25, 30 nt with a custom python script using the transgenes' sense fragments as templates (426 nt for FucT, 314 nt for XylT). We mapped the 19, 21, 25 and 30-nt k-mers against NibSet-1 with BLAT (-minIdentity=100 -tileSize=8 -minMatch=2 -stepSize=1 -minScore=16).

Variant determination based on alignment of assembled contigs against the Nb-1 reference

ΔXT/FT Illumina genomic reads corresponding to 33-fold genomic coverage were assembled with SOAP-denovo2 (-L 500 -K 80 -M 3). The unscaffolded contigs resulting from this assembly were mapped against the Nb-1 draft genome assembly with nucmer (--mum --breaklen=500 --mincluster=60 --maxgap=500 --minmatch=30 --minalign=1000). Nucmer's module "delta-filter" was used to filter alignment results (-i 75.0 -l 1000 -u 10.0). Coordinates from the filtered alignment scores were generated with the "show-coords" module (-T) and were rendered in a non-redundant BED formatted file with bedtools merge. Contigs overlapping coding regions were extracted using bedtools intersect (-u). SNVs were extracted from the filtered alignment scores using the module "show-snps" (-T -r -l) and piped into the mummer2Vcf.pl script (https://github.com/douglasgscofield/bioinfo/blob/master/scripts/mummer2Vcf.pl) to generate a pseudo-VCF formatted file. Variants within coding regions were extracted with bcftools view, providing a non-redundant list of annotated CDS as a condition (-R).

Supplementary Tables

Annotated transposable elements			
Element	Number of elements	Length occupied [bp]	% of sequence occupied
SINEs	85,370	11,479,130	0.46%
LINEs	85,868	68,304,057	2.74%
LTR	1,274,563	1,388,723,157	55.79%
DNA elements	95,760	38,727,508	1.56%
Unclassified	8,782	2,954,159	0.12%
TOTAL	1,550,343	1,510,188,011	60.67%

Supplementary Table S1

	NibSet-1	SGN
Complete Single-copy genes	209	210
Complete Duplicated genes	719	698
Fragmented genes	9	24
% complete + fragmented	98.0%	97.5%
Missing genes	19	24
Total gene groups searched	956	956

Supplementary Table S2

	Number of sequences	Total length [bp]
Nicotiana genus	273,385	113,965,582
Solanaceae family	504,862	205,489,684
A. thaliana	48,315	20,855,795
BLAST Eudicots	1,668,278	744,348,453
BLAST nr protein	106,376,657	38,985,428,197
BLAST nr nucleotide	37,848,925	123,933,400,280

Supplementary Table S3

Gene	Name	Scaffold	Start	End	ΔXT/FT 1 counts	ΔXT/FT 2 counts	WT 1 counts	WT 2 counts
g31184	FucT1	Niben101Scf01272	406	7004	10.83	14.79	21.97	19.23
g80352	FucT-pseudogene	Niben101Scf02631	1984	8357	10.83	12.52	36.03	40.30
g43728	XylT1	Niben101Scf04551	243896	248041	41.15	56.89	96.66	99.83
g40438	XylT2	Niben101Scf04205	352063	356011	24.91	38.68	120.38	103.50

Supplementary Table S4

Gene	Transcript version	Source	Transgene	Match Length	Matches	Mis-matches	Start on transcript	End on transcript
FucT1	EF562630.1	NCBI	FucT	417	415	2	629	1045
	g31184.t1	Augustus	FucT	417	415	2	759	1175
FucT pseudogene	EF562631.1	NCBI	FucT	417	405	12	629	1045
	g80352.t1	Augustus	FucT	417	403	14	1608	1195
XylT1	EF562628.1	NCBI	XylT	310	303	7	1257	1556
	g43728.t1	Augustus	XylT	310	302	8	2117	2426
XylT2	EF562629.1	NCBI	XylT	310	309	1	1240	1549
	g40438.t1	Augustus	XylT	-	-	-	-	-

Supplementary Table S5

Gene	Primer Name	Sequence (5’-3’)
PP2A	Nb_PP2A_Q1F	GACCCTGATGTTGATGTTCGCT
	Nb_PP2A_Q2R	GAGGGATTTGAAGAGAGATTTC
g10744	NbA_Q1F	AATGGTGTTCAGTTTATGGATGC
	NbA_Q2R	TTCAAGAAATACCGGACCAGG
g25290	NbB_Q1F	CAACATTTTCAGGGACACGC
	NbB_Q2R	CAAGCATCCAGTTGTGTCATG
g29021	NbC_Q3F	TGGTACTCCAGGGAAAATGC
	NbC_Q4R	TTCTCACCACCAGGCTTACC
g40387	NbD_Q3F	GCTGGAGTTCCTGCAGATTC
	NbD_Q4R	TGCAAGTTTGTCCACCAAAA
g67787	NbE_Q3F	ATCCAGTTCTTGACGCACCT
	NbE_Q4R	AGTGACGAGGCCATTGAGTT
g76591	NbF_Q3F	ATTATCCTGCTTGGGGGTTC
	NbF_Q4R	TGATACCCTGGATTCCTTGC
g9149	NbG_Q3F	GCGAGACAATGCAACTACGA
	NbG_Q4R	AATTGAAAGGGCCACAGATG
g16390	NbH_Q1F	CCATACCCCTCTCAAGGTCA
	NbH_Q2R	GATAGAGGGATCGCAGCAAC
g21681	NbI_Q1F	GGAGCTTGCACTTTCTTTGG
	NbI_Q2R	CCTTTGCCCACTCTTCTCAG
g29742	NbJ_Q1F	TGGCATTCCCCGTATTCCAT
	NbJ_Q2R	GGCAAATTTCTCCACTGGCA
g45032	NbK_Q3F	GCTGGACAGGATCAGTACGA
	NbK_Q4R	GCTTCCTTTTGTGGCATTGC
g55101	NbL_Q3F	TCATCGAAGAGGAAGTGTCATTTTG
	NbL_Q4R	GATCCTCTGTCCATTCTTCCTGTTTAT
g76921	NbM_Q3F	CACCTCCACCACCTTCGTCCTC
	NbM_Q4R	GCGGTTGCGGCTGTAGTGGTA
g90787	NbN_Q1F	CTTGTGTGAACCCTGAGAGC
	NbN_Q2R	CTTTCGGTTGTGAGGTGCAA

Supplementary Table S6

a

Source of k-mers	Produced k-mers			
	19-mers	21-mers	25-mers	30-mers
FucT-transgene	407	405	401	396
XylT-transgene	295	293	289	284

b

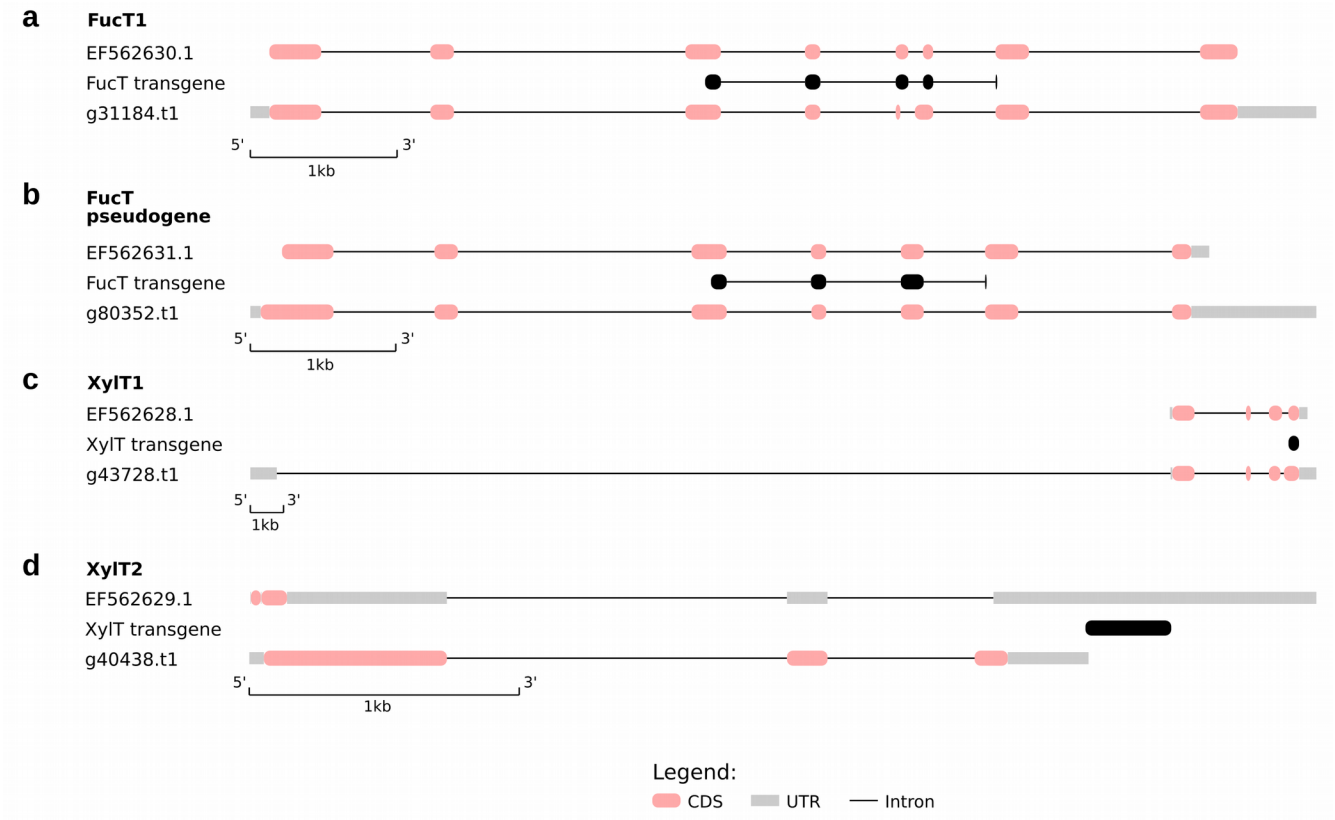
Transcript	Name	Scaffold	Start	End	Mapped k-mers			
					19-mers	21-mers	25-mers	30-mers
g31184.t1	FucT1	Niben101Scf01272	406	7004	370	366	358	348
g80352.t1	FucT-pseudogene	Niben101Scf02631	1984	8357	213	199	171	143
g43728.t1	XylT1	Niben101Scf04551	243896	248041	183	173	153	128
g40438.t1	XylT2	Niben101Scf04205	352063	356011	273	269	261	251
g3481.t1	3-alpha-L-fucosyltransferase A-like	Niben101Scf05494	208147	214675	22	16	8	-
g36277.t1	3-alpha-L-fucosyltransferase A-like	Niben101Scf05447	449846	451560	4	-	-	-

Supplementary Table S7

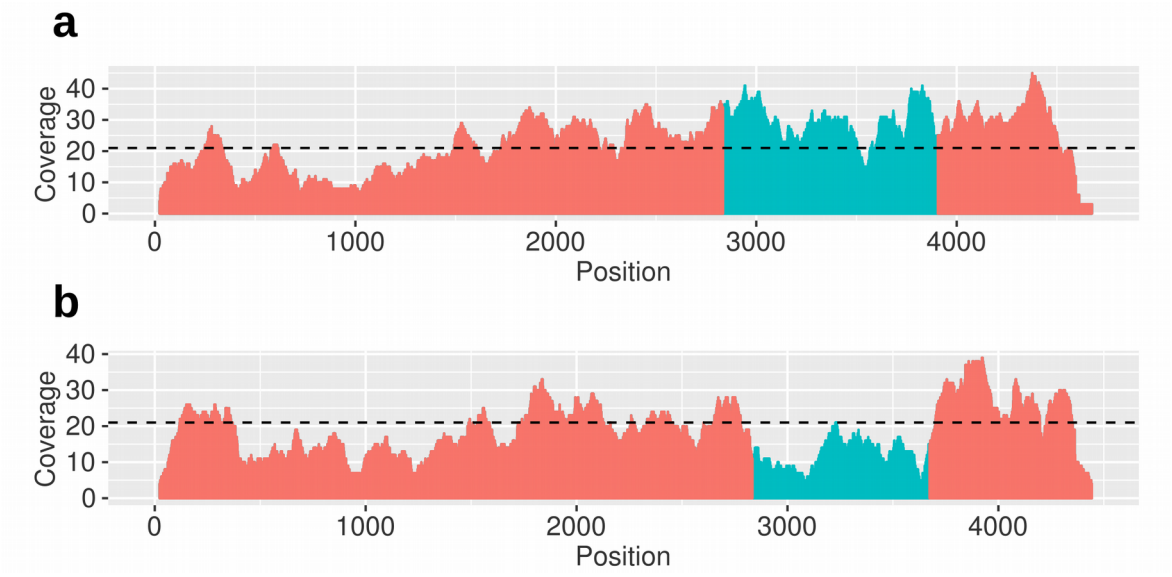
	WT 1	WT 2	ΔXT/FT 1	ΔXT/FT 2
WT 1	-	-	-	-
WT 2	0.989	-	-	-
ΔXT/FT 1	0.997	0.989	-	-
ΔXT/FT 2	0.994	0.977	0.993	-

Supplementary Table S8

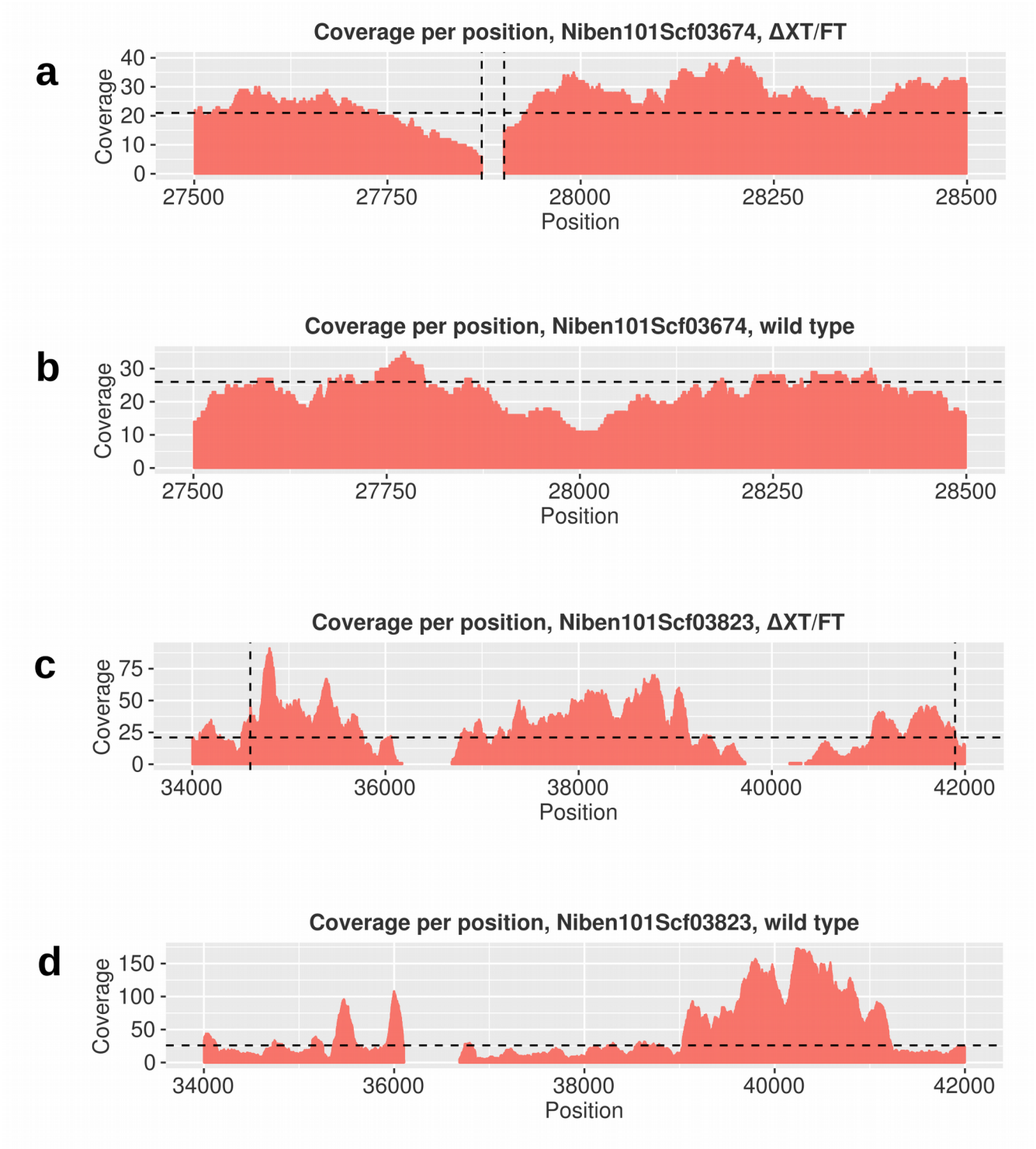
Supplementary Figures



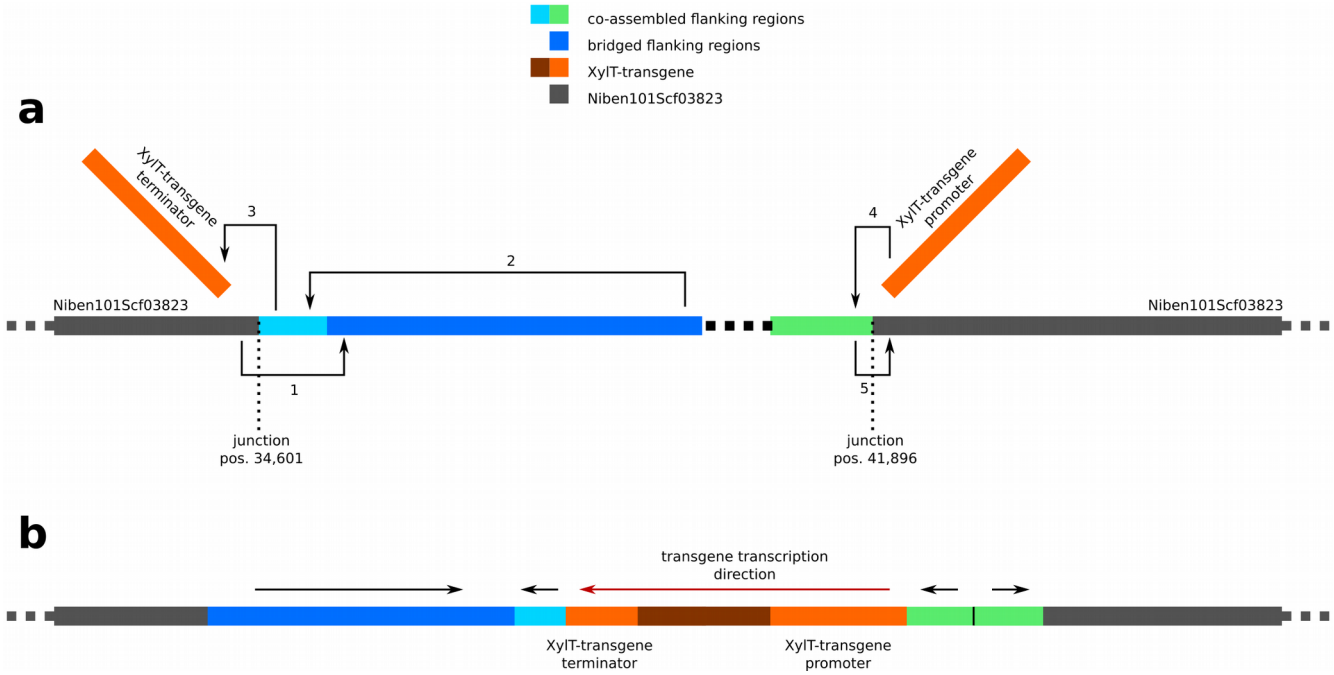
Supplementary Figure S1



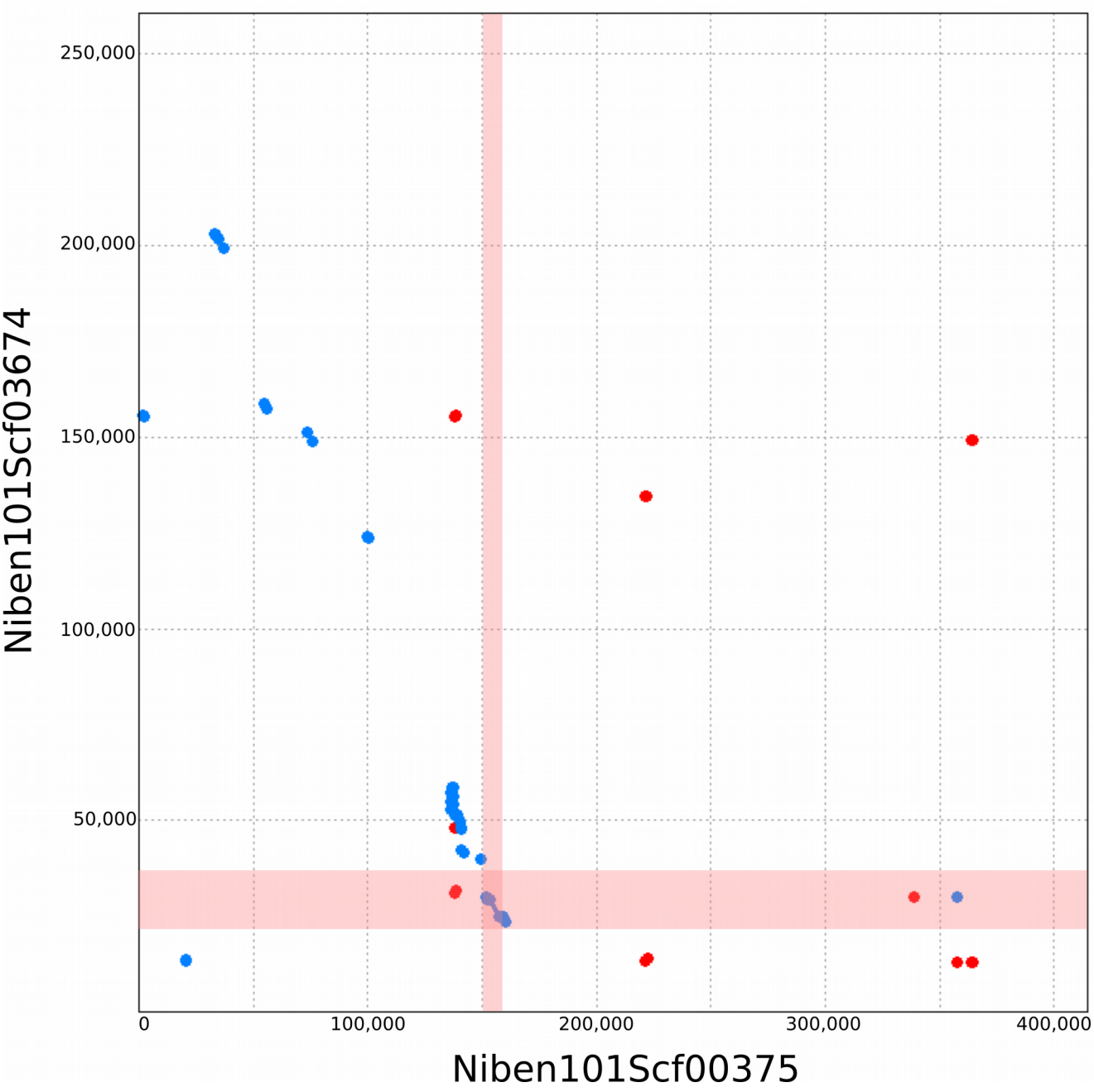
Supplementary Figure S2



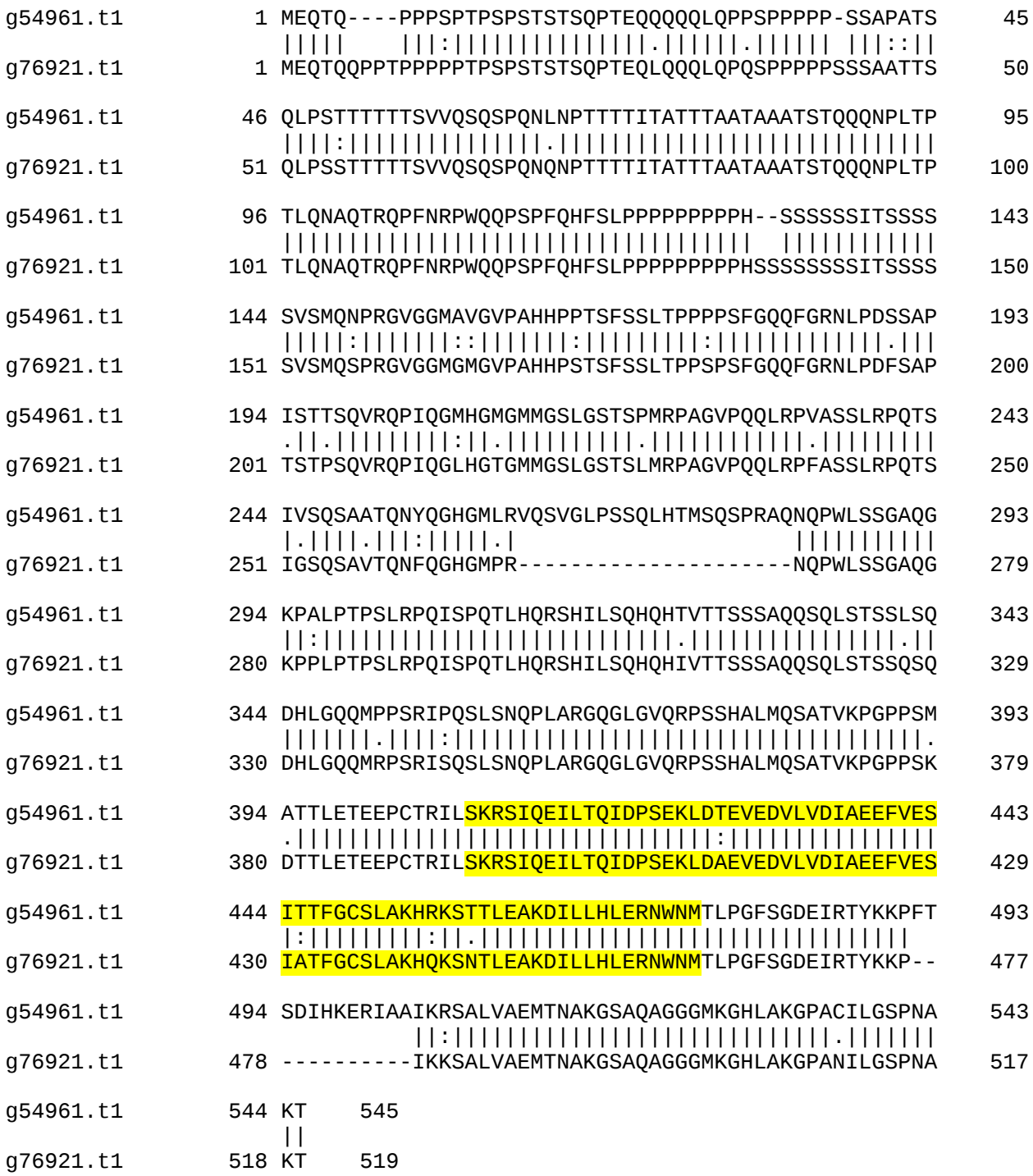
Supplementary Figure S3



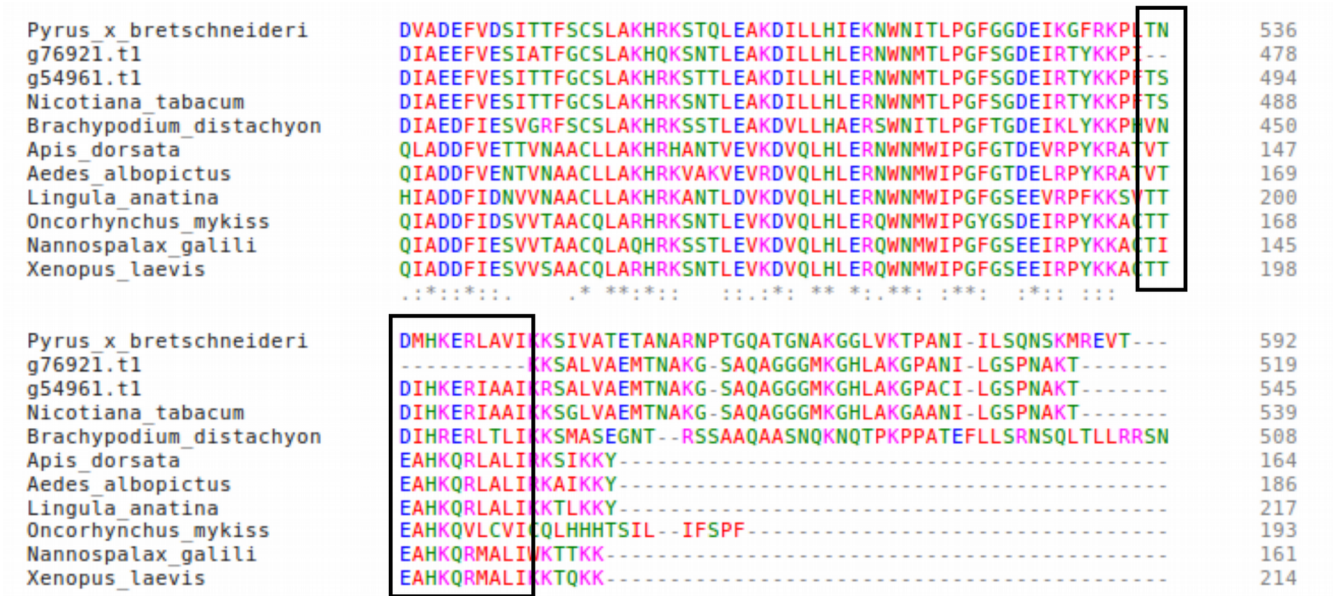
Supplementary Figure S4



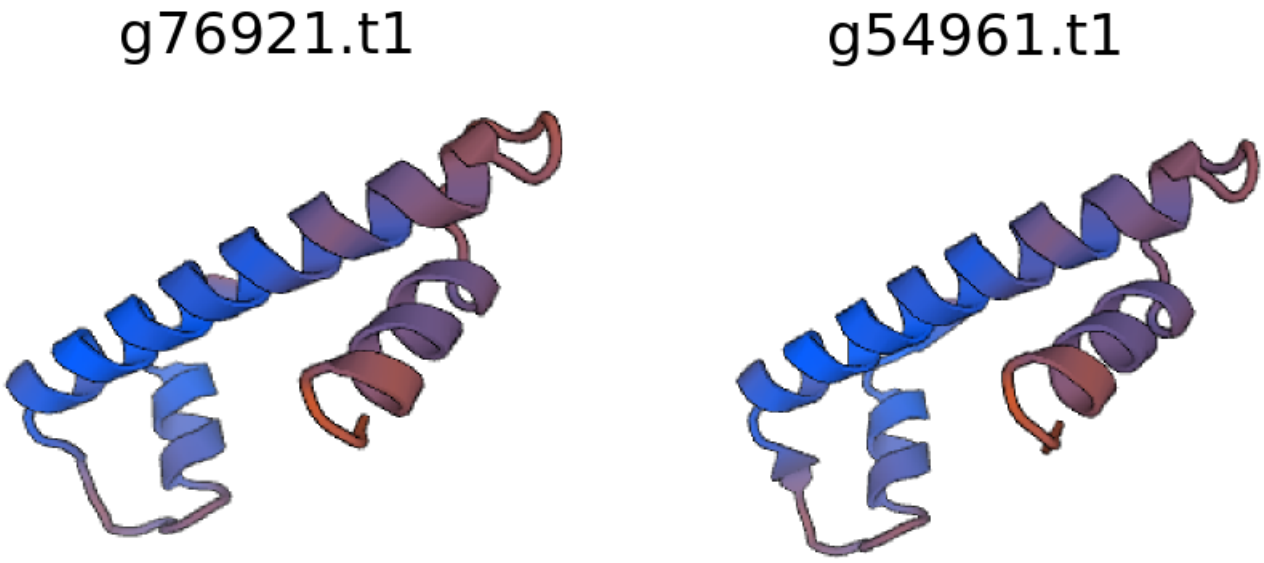
Supplementary Figure S5



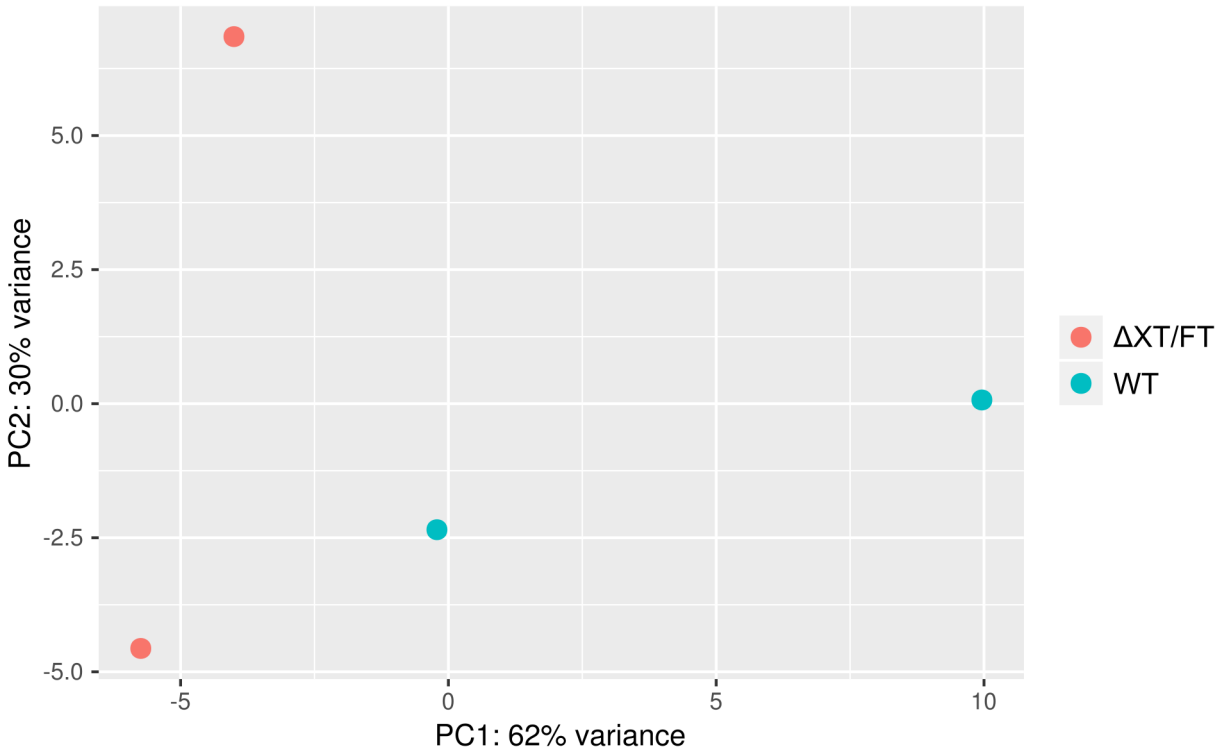
Supplementary Figure S6



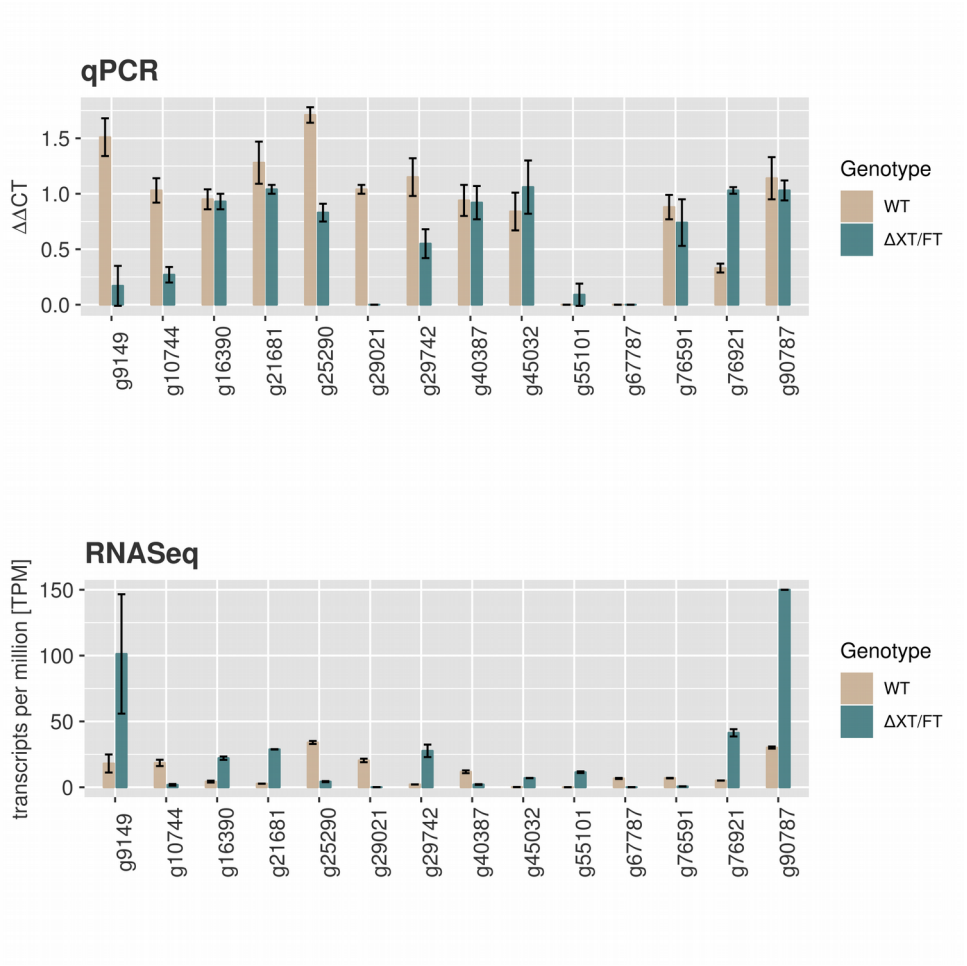
Supplementary Figure S7



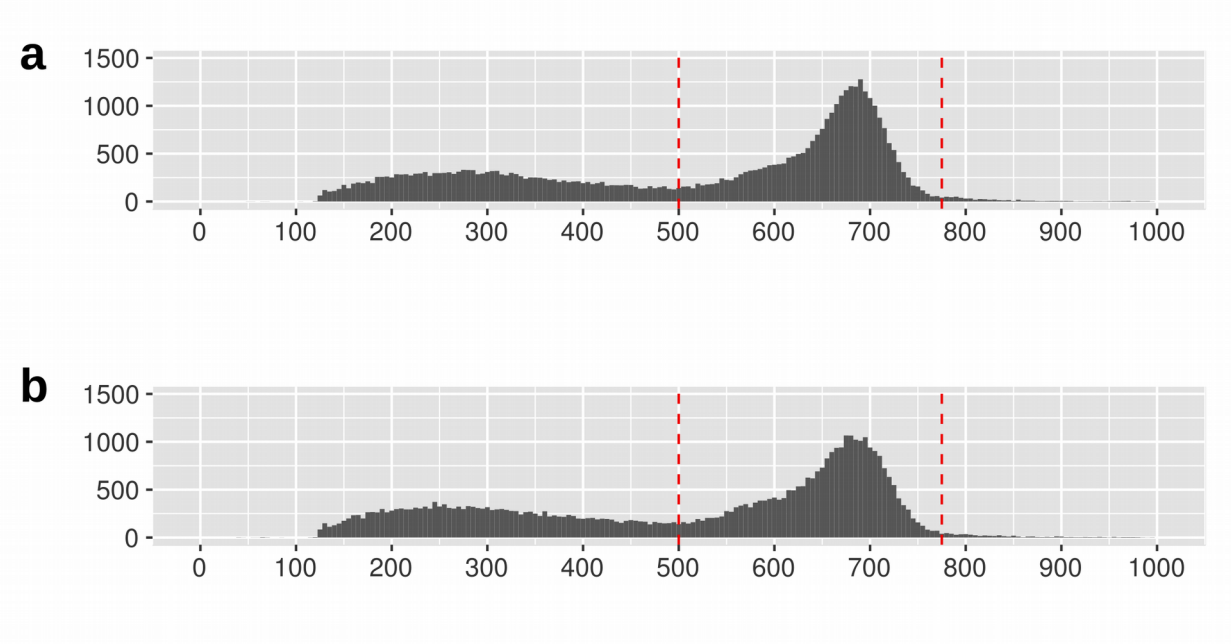
Supplementary Figure S8



Supplementary Figure S9



Supplementary Figure S10



Supplementary Figure S11

Supplementary Table Legends

Supplementary Table S1. Transposable element classes within the *N. benthamiana* reference genome Nb-1 (18). Repeats were identified *de novo* using RepeatModeler; listed here are repeats masked prior to gene prediction.

Supplementary Table S2. Results of a BUSCO analysis to assess completeness of the *N. benthamiana* gene set NibSet-1 and SGN (18).

Supplementary Table S3. Number of sequences and database total length [bp] of each constructed database (downloaded from NCBI in March 2017).

Supplementary Table S4. Normalized counts as computed in DESeq2 (61) for the target genes of the FucT and XylT transgenes.The analysis was performed with wo replicates each for ΔXT/FT (light yellow) and for the wild type (dark yellow).

Supplementary Table S5. Regions of the FucT1, FucT-pseudogene, XylT1 and XylT2 transcripts (both in the NCBI and the NibSet-1 version) that are targeted by the transgenes, as obtained with Blast (see methods). The transcript version reports either the accession number of Genbank (source: “NCBI”) or the gene name in NibSet-1 (source: “Augustus”). Reported are: the targeting transgene, the match length (nt), the number of matches (nt), the number of mismatches and the start-end coordinates on the transcript sequences. We report also a non-match for g40438.t1 as a means of comparison with its NCBI counterpart. Reasons are discussed within the supplementary text.

Supplementary Table S6. List of primer sequences used to perform qPCR on the differentially expressed genes found with mRNA-seq.

Supplementary Table S7. K-mer generation from the transgene sequences and their mapping onto NibSet-1. For the k-mer generation, only the sense fragments of the transgenes were used (426 nt for the FucT-transgene, 314 nt for the XylT-transgene). **a)** Total number of possible k-mers that can be generated from the transgene sequences. **b)** Number of k-mers mapped to each target (yellow) and off-target (white), for each tested k-mer size (19, 21, 25, 30). A dash represents absence of mapped k-mers.

Supplementary Figure Legends

Supplementary Figure S1. Each panel, from top to bottom: gene models obtained by mapping the NCBI FASTA sequences of the FucT and XylT genes onto the Nb-1 draft genome assembly; region targeted by the corresponding transgene; gene model predicted in NibSet-1. Gray: untranslated regions (UTR); red: coding regions (CDS); black blocks: transgene; thin lines: introns. All transcripts are shown from 5' (left) to 3' (right) regardless of the orientation in the genome. For each panel, a 1kb scale is shown. **a)** FucT1 (EF562630.1, g31184.t1), **b)** FucT-pseudogene (EF562631.1, g80352.t1), **c)** XylT1 (EF562628.1, g43728.t1), **d)** XylT2 (EF562629.1, g40438.t1).

Supplementary Figure S2. Genomic coverage per position in each of the two transgenes present within the ΔXT/FT genome. As both transgenes shared the same promoter and the same terminator region, in red we show the coverage in shared regions, while in cyan we show the coverage in the transgene-specific region. Horizontal dashed lines represent the average observed genomic coverage (~21x). **a)** FucT-transgene. **b)** XylT-transgene.

Supplementary Figure S3. Genomic coverage per position (red) in ΔXT/FT (**a** and **c**) and wild type (**b** and **d**) on scaffold Niben101Scf03674 (**a** and **b**) and Niben101Scf03823 (**c** and **d**). Horizontal dashed lines represents the average observed genomic coverage (~21x in ΔXT/FT, ~26x in the wild type). Vertical dashed lines show the junctions identified by chimeric reads (see results). Such junctions are visible only in the ΔXT/FT, as no transgenes are present in the wild type.

Supplementary Figure S4. Results of a re-assembly of the region where the insertion of the XylT transgene took place in scaffold Niben101Scf03823. **a)** Green and blue segments represent reassembled scaffolds in their mapping locations on the Nb-1 assembly (18). Orange segments show the transgene promoter and terminator. Numbered arrows indicate the ordering of the connections. Vertical dashed lines indicate the junctions, with their position. The horizontal dashed segment shows the uncovered region described in the results (positions 39,727 to 40,181). Grey segments indicate regions outside of the insertion junctions in scaffold Niben101Scf03823. **b)** Reordering of the insertion region, according to bridging pairs. Arrows indicate the proposed direction of each segment. The red arrow indicates both the direction of the transgene in the genome and its transcription direction. Note: the green segment is repeated twice, with reversed orientation.

Supplementary Figure S5. Alignment between the two scaffolds that contain genes g76921 and

g54961, showing aligned regions on the forward strand (red) and on the reverse strand (blue). The intersection of the highlighted areas indicates the positions where g76921 and g54961 map (scaffold Niben101Scf03674: gene g76921, positions 23406-38910; scaffold Niben101Scf00375: gene g54961 positions 151407-160415). The blue dots are approximately located on a diagonal, indicating that these two scaffolds are most likely derived from homeologous chromosomes that were inherited from the two ancestral species which gave rise to allotetraploid *N. benthamiana*.

Supplementary Figure S6. Global protein sequence alignment between genes g76921 and g54961, both annotated as TFIID subunit 12-like isoform X1. Pipes (|) indicate identical amino acids, colons (:) indicate different amino acids that share chemical properties (scoring > 0.5 in the Gonnet PAM 250 matrix), dots (.) indicate different amino acids that do not have similar chemical properties (scoring =< 0.5 in the Gonnet PAM 250 matrix). The region defining the TFIID 20 kDa protein family (PF03847) is marked in yellow in both sequences.

Supplementary Figure S7. Section of a multiple sequence alignment performed with g76921.t1, g54961.t1 and further protein sequences annotated as TFIID subunit 12-like isoform X1 available in NCBI-Protein. Black boxes indicate a region that only g76921.t1 is lacking.

Supplementary Figure S8. Folding of *N. benthamiana* proteins encoded by genes g76921 and g54961, as predicted with SWISS-MODEL.

Supplementary Figure S9. Principal component analysis (PCA) performed on the normalized read counts obtained from each replicate in each condition (total: four data points). Colors indicate the condition which each replicate (i.e. dot) belongs to.

Supplementary Figure S10. a) ΔΔCT values detected through qPCR in ΔXT/FT and WT (mean of three replicates), for each identified potential differentially expressed gene (DEG). Black lines indicate standard deviations among replicates. **b)** Transcripts per million (TPM) detected through mRNA-seq in ΔXT/FT and WT (mean of two replicates), for each identified potential DEG.

Supplementary Figure S11. Insert size estimation of the ΔXT/FT and WT genomic sequencing libraries, based on 50,000 read-pairs each mapped against the Nb-1 draft genome assembly (18). Red vertical dashed lines indicate the boundaries of insert size that were chosen for mapping (left: 500 bp; right: 775 bp). **a)** ΔXT/FT **b)** WT.

Supplementary References

70. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research. 2016 Jan 4;44(D1):D279–85.

71. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974 Jan 15;13(2):222–45.

72. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research. 2014 Jul 1;42(W1):W252–8.

73. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 2003 Jul 1;31(13):3381–5.

74. Deng Y, Wang CC, Choy KW, Du Q, Chen J, Wang Q, et al. Therapeutic potentials of gene silencing by RNA interference: Principles, challenges, and new strategies. Gene. 2014 Apr;538(2):217–27.

75. Martinez J, Patkaniowska A, Urlaub H, Lührmann R, Tuschl T. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell. 2002 Sep 6;110(5):563–74.

76. Eulalio A, Huntzinger E, Izaurralde E. Getting to the Root of miRNA-Mediated Gene Silencing. Cell. 2008 Jan;132(1):9–14.

77. Thomas CL, Jones L, Baulcombe DC, Maule AJ. Size constraints for targeting post-transcriptional gene silencing and for RNA-directed methylation in Nicotiana benthamiana using a potato virus X vector: Size constraints for mediating PTGS and transgene methylation. The Plant Journal. 2001 Dec 23;25(4):417–25.

78. Baulcombe D. RNA silencing. Current Biology. 2002 Feb;12(3):R82–4.

79. Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. Nature. 2004 Sep 16;431(7006):343–9.

80. Senthil-Kumar M, Hema R, Anand A, Kang L, Udayakumar M, Mysore KS. A systematic study to determine the extent of gene silencing in Nicotiana benthamiana and other Solanaceae species when heterologous gene sequences are used for virus-induced gene silencing. New Phytol. 2007;176(4):782–91.

81. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.

82. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience [Internet]. 2012 Dec [cited 2018 Feb 6];1(1). Available from: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/2047-

83. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16(6):276–7.

84. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004 Aug 19;5:113.

85. Eyrich VA, Przybylski D, Koh IYY, Grana O, Pazos F, Valencia A, et al. CAFASP3 in the spotlight of EVA. Proteins: Structure, Function, and Genetics. 2003;53(S6):548–60.

Chapter 3:

Parental origin of the allotetraploid tobacco *Nicotiana benthamiana*

Published as the article:

Schiavinato, M., Marcet-Houben, M., Dohm, J.C., Gabaldón, T. and Himmelbauer, H. (2019), Parental origin of the allotetraploid tobacco *Nicotiana benthamiana*. *Plant J.* Accepted Author Manuscript. doi:10.1111/tpj.14648

Parental origin of the allotetraploid tobacco *Nicotiana benthamiana*

Matteo Schiavinato[1], Marina Marcet-Houben[2,3], Juliane C. Dohm[1], Toni Gabaldón[2,3,4], Heinz Himmelbauer[1]*

[1]Institute of Computational Biology, Department of Biotechnology, University of Natural Resources and Life Sciences (BOKU), Muthgasse 18, 1190 Vienna, Austria
[2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
[3]Current address: Barcelona Supercomputing Centre (BSC-CNS) and Institute for Research in Biomedicine (IRB), Barcelona, Spain
[4]ICREA, Barcelona, Spain

*To whom correspondence should be addressed: Heinz.Himmelbauer@boku.ac.at

Running title: *N. benthamiana* phylome and subgenome separation

Key words: *Nicotiana benthamiana*, tobacco, interspecific hybrid, subgenome separation, phylome analysis, plant genomics, phylogenetic tree

Summary

Nicotiana section *Suaveolentes* is an almost all-Australian clade of allopolyploid tobacco species including the important plant model *Nicotiana benthamiana.* The homology relationships of this clade and its formation are not completely understood. To address this gap, we assessed phylogenies of all individual genes of *N. benthamiana* and the well-studied *N. tabacum* (section *Nicotiana*) and their homologs in six diploid *Nicotiana* species. We generated sets of 44,424 and 65,457 phylogenetic trees of *N. benthamiana* and *N. tabacum* genes, respectively, each collectively called a phylome. Members of *Nicotiana* sections *Noctiflorae* and *Sylvestres* were represented as the species closest to *N. benthamiana* in most of the gene trees. Analysing the gene trees of the phylome we 1) dated the hybridization event giving rise to *N. benthamiana* to 4-5 MyA, and 2) separated the subgenomes. We assigned 1.42 Gbp of the genome sequence to section *Noctiflorae* and 0.97 Gbp to section *Sylvestres* based on phylome analysis. In contrast, read mapping of the donor species did not succeed for separating the subgenomes of *N.*

benthamiana. We show that the maternal progenitor of *N. benthamiana* was a member of section *Noctiflorae*, and confirm a member of section *Sylvestres* as paternal subgenome donor. We also demonstrate that the advanced stage of long-term genome diploidization in *N. benthamiana* is reflected in its subgenome organisation. Taken together, our results underscore the usefulness of phylome analysis for subgenome characterization in hybrid species.

Introduction

Nicotiana benthamiana is an Australian tobacco species, mostly popular as platform for recombinant protein production (van Herpen *et al.*, 2010; Bally *et al.*, 2018). The *Nicotiana* genus is part of the *Solanaceae*, a family that includes many economically relevant plants such as tomato, potato and eggplant. The genus *Nicotiana* is organized into several sections, five of which contain polyploids formed by interspecific hybridization (Figure **1**) (Knapp *et al.*, 2004; Leitch *et al.*, 2008). Among these, *N. benthamiana* belongs to the section *Suaveolentes*, an almost all-Australian clade with the exception of *N. africana*.

It has repeatedly been shown that the paternal progenitor in the hybridization that resulted in the species of section *Suaveolentes* is affiliated with the section *Sylvestres (Knapp et al.*, 2004; Leitch *et al.*, 2008; Clarkson *et al.*, 2010; Wang and Bennetzen, 2015; Bally *et al.*, 2018). The origin of the maternal progenitor is less clear, and different suggestions have been made depending on which phylogenetic markers were used (Leitch *et al.*, 2008; Kelly *et al.*, 2012; Clarkson *et al.*, 2017). A certain preference for section *Noctiflorae* has been consistent throughout the literature, mostly based on analyses involving multiple regions of the *N. benthamiana* plastid genome which is maternally inherited (Aoki and Ito, 2000; Clarkson *et al.*, 2004). However, none of the previous studies could yet clearly resolve the maternal phylogeny.

The date of the hybridization event has been estimated multiple times. It was believed that section *Suaveolentes* originated more than 10 million years ago (MyA) (Clarkson *et al.*, 2004; Leitch *et al.*, 2008; Kelly *et al.*, 2012; Bally *et al.*, 2018). However, another study dated the hybridization event at 6 MyA, followed by a lag phase whereby the original hybrid expanded throughout the Australian territory, favored by a humid climate, and during the Pleistocene the aridification of Australia coincided with a radiation of the clade that has led to more than 70 extant *Suaveolentes* species (Clarkson *et al.*, 2017). The positive response towards aridification is a rare and qualifying property, attributed to the uncommon genome plasticity of species from this section (Bally *et al.*, 2015).

Nicotiana hybrid species show a high level of intermixing between their subgenomes (Lim *et al.*, 2007). Tobacco species affiliated with section *Suaveolentes* likely are the oldest *Nicotiana* polyploids (Clarkson *et al.*, 2004) in an advanced stage of long-term genome diploidization. Many members of this section have fewer than n=24 chromosomes, which is less than the sum of the chromosomes of their parents (n=12) (Leitch *et al.*, 2008). A high genetic turnover has been shown in some *Nicotiana* hybrids (Lim *et al.*, 2007), involving intense mobilization of retrotransposons

(Petit *et al.*, 2010) and genome downsizing mostly affecting the paternally derived subgenome (Renny-Byfield *et al.*, 2011). *N. benthamiana* is therefore expected to have undergone extensive rearrangements among its two subgenomes.

Our study aims at addressing the open questions on the maternal progenitor of *N. benthamiana* and the time of the interspecific hybridization event that gave rise to *N. benthamiana*. We address both questions using a high-throughput phylogenetic approach involving the reconstruction of phylogenetic trees for all genes encoded in a genome, referred to as a phylome (Huerta-Cepas *et al.*, 2007). By inspecting the topologies of individual gene trees, we infer the most likely pair of parents for *N. benthamiana* and use this information to separate its subgenomes. We validate our strategy by comparisons to *Nicotiana tabacum*, for which the details of the hybridization are known (Sierro *et al.*, 2014; Edwards *et al.*, 2017).

Results

Selection and preparation of Nicotiana gene sets

All the polyploid species in the *Nicotiana* genus originated from diploid *Nicotiana* parents (Clarkson *et al.*, 2004; Kelly *et al.*, 2012). Hence, we studied the parental origin of *N. benthamiana* through its homology relationships with diploid *Nicotiana* species each representing a different section. We included four publicly available gene sets from diploid tobacco species in our analysis, i.e. from *N. attenuata* and *N. obtusifolia (Xu et al.*, 2017), as well as from *N. sylvestris* and *N. tomentosiformis* (Sierro *et al.*, 2013). Gene sets from representatives of two additional sections, *N. cordifolia* and *N. noctiflora*, were assembled by us using previously generated mRNA-Seq data from leaf tissue (Long *et al.*, 2016). After quality-trimming we assembled 40 million read-pairs for *N. cordifolia* and 89 million read-pairs for *N. noctiflora* into an initial set of 147,902 *N. cordifolia* transcripts (avg. length 1127 nt) and 168,632 *N. noctiflora* transcripts (avg. length 1026 nt). The final gene sets contained 22,251 genes for *N. cordifolia* (avg. length 1480 nt) and 22,940 genes for *N. noctiflora* (avg. length 1494 nt) (Table **1**). We assessed the completeness by homology searches to 1440 highly conserved plant genes using BUSCO (Simão *et al.*, 2015) of which 75% were matched in the assembled transcriptome of *N. cordifolia* and 78% were matched in the assembled *N. noctiflora* transcriptome. These values, which are below a fully satisfactory BUSCO score, can be explained by the absence of tissue diversity as the *N. cordifolia* and *N. noctiflora* transcript sequences were exclusively derived from leaf samples. Nevertheless, we considered that these *de novo* assembled gene sets would provide a reasonably good representation of the transcriptomes of *N. cordifolia* and *N. noctiflora*.

Additionally, we included the gene sets from *N. benthamiana (Schiavinato et al.*, 2019) and from *N. tabacum (Edwards et al.*, 2017) comprising 50,503 and 69,500 protein-coding genes, respectively. *N. tabacum* was used as control throughout our study as it is a well-characterized hybrid with known parents (namely *N. sylvestris* and *N. tomentosiformis*); the hybridization event has been

dated between 0.2 and 0.4 MyA (Sierro *et al.*, 2014; Edwards *et al.*, 2017). Our final collection featured six diploid *Nicotiana* species (*N. attenuata, N. cordifolia, N. noctiflora, N. obtusifolia, N. sylvestris* and *N. tomentosiformis*) and two allotetraploid species (*N. benthamiana* and *N. tabacum*) (Table **1**).

Phylome reconstruction

Using the PhylomeDB pipeline (Huerta-Cepas *et al.*, 2014) we built a phylome for *N. benthamiana* and one for *N. tabacum* using their gene sets as seeds to study homology relationships between their genes and those of the six diploid *Nicotiana* species (Table **1**). PhylomeDB requires at least three homologs within the data set (orthologs or paralogs), to build a gene tree, a condition that was true for 44,424 (88.7%) *N. benthamiana* seed genes, representing the backbone phylome. The remaining 5,666 genes did not return a tree for various reasons: 281 genes had no homologs, 3,390 had only one or two homologs, and 1,995 genes showed premature stop codons within the CDS. The latter can be attributed to result from technical issues, such as problems of sequencing data quality, or artefacts that arose during *de novo* genome assembly or gene prediction, but could also be due to biological reasons, such as gene birth and decay (Nei and Rooney, 2005). The addition of the assembled transcripts from *N. cordifolia* and *N. noctiflora* extended 32,614 trees by at least one leaf. A total of 29,330 trees in the *N. benthamiana* phylome featured at least five species and passed all filtering criteria (see methods): these trees were used to detect the parental progenitor and to date the hybridization event. The *N. tabacum* phylome returned a tree for 65,457 seed sequences (94.4% of genes), 47,832 of which passed all our filtering criteria.

Phylogenetic distance between taxa

To determine the closest relative of either *N. benthamiana* or *N. tabacum* in each gene tree, we studied the gene or gene subtree that was placed in the tree partition closest to the seed gene and belonged to a species different from the seed, using its taxonomic origin as to infer parental relationships. In 20,947 *N. benthamiana* trees and 36,239 *N. tabacum* trees these “closest sister leaves” (CSLs) were not sufficiently supported, too noisy or phylogenetically too distant, and were discarded (Figure **2**, see methods). The remaining 8,383 and 11,593 gene trees of *N. benthamiana* and *N. tabacum*, respectively, revealed different occurrences of each *Nicotiana* section as CSL whereby the two most recurring sections can be considered as closest relatives, i.e. as parental sections. In most cases the assignment to a section was straightforward as only one section was present in the closest sister subtree (5,565 and 9,889 gene trees in *N. benthamiana* and *N. tabacum*, respectively). In the other cases the section of the leaf (i.e. terminal node) that showed the shortest distance to the seed gene (as inferred from the sum of the branch lengths connecting the relevant nodes) was considered. In the *N. benthamiana* gene trees, the highest occurrences as CSLs were found for sections *Noctiflorae* and *Sylvestres* (Figure **3a**). In the *N. tabacum* gene trees, sections *Sylvestres* and *Tomentosae* showed the highest occurrences as CSLs, supporting

previous findings (Figure **3b**). This result was confirmed when analysing the distribution of phylogenetic distances between the seed and the closest leaf from each taxon (Figure **3c**). For *N. benthamiana* we identified four groups of sections, ordered by distance: (1) *Sylvestres*; (2) *Noctiflorae* and *Petunioides;* (3) *Tomentosae*; (4) *Paniculatae* and *Trigonophyllae*. In the case of *N. tabacum*, the two sections *Sylvestres* and *Tomentosae*, most likely the two parental sections as described previously, were clearly separated from all other sections (Figure **3d**). Based on these data we considered sections *Noctiflorae* and *Sylvestres* as the most likely parental sections for *N. benthamiana*.

Cluster network

Homeologous genes from different subgenomes often produce phylogenetic trees that show incongruent topologies; such incongruencies are a strong indicator of hybridization. Once we identified the two mostly likely parental sections for *N. benthamiana* and *N. tabacum*, we assessed whether we could observe such a hybridization with a cluster network (Huson and Rupp, 2008). We isolated phylogenetic trees whereby the CSL belonged to one of the top three sections most commonly found as CSLs, that is, sections *Noctiflorae, Sylvestres* and *Trigonophyllae* for *N. benthamiana* (Figure **3a**), and sections *Sylvestres, Trigonophyllae* and *Tomentosae* for *N. tabacum* (Figure **3b**). Section *Trigonophyllae* was included as it was the one found most often as an alternative candidate parent throughout our analysis. From each of these two pools we created ten random subsets of 100 trees, and we generated a cluster network from each subset. A single hybridization signal between sections *Noctiflorae* and *Sylvestres* was detected in 8/10 networks generated for *N. benthamiana*. For *N. tabacum*, a single signal was observed between sections *Sylvestres* and *Tomentosae* in 10/10 networks. Hence, we confirmed that sections *Noctiflorae* and *Sylvestres* are the most likely parental sections for *N. benthamiana*. We report the most recurrent topology for each of the two species as a circular cladogram (Figure **4**).

N. benthamiana subgenome analysis using transcriptome data

As a different means of evaluating the origin of the *N. benthamiana* subgenomes we mapped Illumina mRNA-Seq data from different species of tobacco against the *N. benthamiana* genome. For each diploid species of tobacco in the *N. benthamiana* phylome, we used a subset of eight million quality-trimmed read-pairs available from publicly available transcriptomic data (Table **1**) (Sierro *et al.*, 2013; Long *et al.*, 2016). After mapping, we calculated the coverage per position on 81,412,192 positions within annotated exons in the *N. benthamiana* genome based on a gene set that we had calculated previously (named "NibSet-1") (Schiavinato *et al.*, 2019).. Our results show that transcriptomic reads from species of the sections *Noctiflorae* and *Sylvestres* covered more positions than the other species (Figure **5a**). We then determined which pair of candidate parents covered the NibSet-1 gene set most extensively. To do so, we combined the results from the known paternal subgenome donor section *Sylvestres* with those of each other diploid species. We

assessed what proportion of the NibSet-1 exome was uniquely covered by either one of the two candidates, and by both. Again, we observed that a combination of section *Sylvestres* with section *Noctiflorae* makes the best parental pair (Figure **5b**), closely followed by the pair formed with section *Trigonophyllae*; section *Petunioides* returned the lowest result.

Subgenome intermixing

Genomic data of one of the parental sections may be used to identify the portion of the genome that originated from this parental section by mapping genomic reads onto the hybridized genome. In case the two subgenomes were still distinguishable and each assignable to one of the parents this should be reflected in the mapping result, i.e. one fraction of the genome would be covered well whereas the other fraction remains sparsely covered. However, if the subgenomes were highly intermixed a clear separation would probably be impossible. To assess the level of intermixing between subgenomes we mapped genomic paired-end reads from species affiliated with section *Noctiflorae* (*N. glauca*) and *Sylvestres* (*N. sylvestris*) onto the Nb-1 draft genome assembly, allowing for up to 5% divergence between reads and reference. Although only < 20% of the reads matched sequences of the Nb-1 assembly, the total number of matched scaffolds comprised 95% of Nb-1. Regardless of the species from which the reads came from, these scaffolds were only partially covered by reads (min. coverage: 1x). *N. glauca* reads covered around 12% of their length (mean: 11.9%, median: 11.0%, max: 78.0%), while *N. sylvestris* reads covered around 13% of it (mean: 13.0%, median: 11.9%, max: 78.9%) (Figure **6a**). When considering only coding regions (CDS), *N. glauca* reads coverage increased to 60% (mean: 60.2%, median: 63.1%, max: 100.0%), while *N. sylvestris* to 65% (mean: 65.4%, median: 69.0%, max: 100.0%) (Figure **6b**). In both cases, it was not possible to distinguish two separate groups of scaffolds that would represent the two subgenomes.

For comparison, we mapped genomic reads from species affiliated with sections *Sylvestres* (*N. sylvestris*) and *Tomentosae* (*N. tomentosiformis*) against the scaffolds of the *N. tabacum* Nitab-v4.5 assembly (Edwards *et al.*, 2017), allowing a divergence of 2%, and computed the covered fraction in the same way as for *N. benthamiana*. This time, 84.0% of the *N. sylvestris* reads and 72.4% of the *N. tomentosiformis* reads were mapped, and we could clearly identify two groups of scaffolds, i.e. one highly covered fraction (range: 50-100%) and one poorly covered fraction (range: 0-50%) (Figure **6c**). Within the highly covered fraction, the *N. sylvestris* reads covered around 95% of the scaffold lengths (mean: 95.4%, median: 98.0%, max: 100.0%), while the *N. tomentosiformis* covered around 89% (mean: 88.8%, median: 93.4%, max: 99.8%). Within the poorly covered fraction, the *N. sylvestris* reads covered around 2% (mean: 1.5%, median: 0.3%, max: 49.5%), and the same for *N. tomentosiformis* reads (mean: 2.3%, median: 0.7%, max: 50.0%). Within CDS regions the distinction between the two groups was even sharper (Figure **6d**). The highly covered fraction approximated a full coverage for both *N. sylvestris* (mean: 99.1%, median: 100.0%, max: 100.0%) and *N. tomentosiformis* (mean: 96.0%, median: 99.4%, max: 100.0%), while the poorly

covered fraction maintained low coverage (mean: 3.4%, median: 1.6%, max: 49.7% for *N. sylvestris*; mean: 7.1%, median: 5.0%, max: 49.4% for *N. tomentosiformis*). This result confirmed on the one hand that mapping of genomic reads can identify separate subgenomes and suggested on the other hand that the subgenomes of *N. benthamiana* were much more intermixed than those of *N. tabacum*.

Subgenome separation

Since genomic mapping was shown to be hardly promising for the separation of highly intermixed subgenomes we attempted to separate the two subgenomes of *N. benthamiana* using the phylogenetic information provided within its phylome. By considering only gene-encoding regions we ruled out the high repetitive content and reduced the effect of the advanced genomic intermixing state of the two subgenomes. Out of 44,424 trees, we extracted 43,597 trees whereby each sequence in the underlying multiple sequence alignment showed at most 10% undetermined characters ('N'). In each tree we extracted the leaf belonging to a parental section that was closest to the *N. benthamiana* seed sequence. Such leaves were considered as reliable when they showed a node support of at least 0.9. A total of 18,071 extracted leaves belonged to section *Noctiflorae* and 15,511 to section *Sylvestres*. We refer to a seed gene as “orphan” if no leaf of a parental section was contained in its tree (10,015 trees). Wherever possible we inferred the parental information for orphan trees from neighboring genes in the genome sequence (see methods). A maternal origin (*Noctiflorae*) was inferred for 1,881 orphan genes, while a paternal origin (*Sylvestres*) was inferred for 1,501 orphan genes. Our final set of assigned genes comprised 19,952 maternal genes and 17,012 paternal genes; 6,633 genes remained orphans.

We used the genes' parental information to assign a parental origin to the Nb-1 assembly scaffolds. Scaffolds containing at least one gene sum up to 95% of the total assembly size (10,363 scaffolds, mean 4.9 genes per scaffold). Scaffolds containing genes from both parents were only assigned if one parent was represented by at least 75% of the genes. Genes of unknown parental origin were ignored. In this way, we assigned a maternal origin to 1.42 Gbp (50.3%) and a paternal origin to 0.97 Gbp (34.4%) of Nb-1 (Table **2**, Supplementary Files **1** and **2**).

Dating the hybridization event

To date the hybridization events that had led to the formation of *N. benthamiana* and *N. tabacum,* respectively*,* we studied fourfold-degenerate (4D) sites (Lagerkvist, 1978). 4D sites are third positions in codons that can mutate to any nucleotide without changing the translated amino acid (Topal and Fresco, 1976). Hence, they represent an ideal molecular clock in a situation where few calibration points are available. First, we assessed the transversion (Tv) ratio at these sites, computing the 4DTv ratio in each tree with a valid CSL, that is, 8,383 trees in *N. benthamiana* and 11,593 trees in *N. tabacum*. A 4DTv ratio is only a relative measurement of divergence time, i.e. it cannot be converted into an absolute time measurement such as years. However, concordance

between the relative divergence times of the parental sections could corroborate our claim on the parental section choice. Within each tree we computed the 4DTv ratio only between the seed sequence and its CSL. We applied the criteria that, firstly, an alignment needed to encompass at least five substitutions to compute a ratio, and, secondly, that seed and CSL must diverge by at most 0.05 substitutions per site. We chose such strict filtering thresholds in order to exclude genes which either evolve particularly slow or fast. We plotted the 4DTv ratios obtained from each tree to analyse their distribution. Ideally, a species 4DTv ratio is defined by a sharp peak in the distribution, meaning that most of its gene trees returned similar 4DTv ratios. Since they define relative time distances, the ideal case would be that only the two parental species have a peak at the same 4DTv ratio, confirming that they diverged at the same time from the hybrid. We were not able to distinguish *N. benthamiana*'s parents based on this criterion: the peaks obtained from most sections overlapped to a great extent, with the exception of section *Tomentosae* and *Trigonophyllae* (Figure **7a**). The results obtained within the *N. tabacum* phylome instead showed a strong peak concordance only for the two parental sections (Figure **7b**).

We then used the same sets of trees to obtain an absolute time distance between the hybrid and the parental species (i.e. a hybridization date). We used the substitutions per position at fourfold degenerate sites (this time regardless of the nature of the substitution) as a measure of sequence divergence under neutral evolution. In each tree the value was computed only between the seed and its CSL. We divided those values by a rate of substitutions per position per generation time (5e-09) in the range of those expected for plant nuclear genes (Wolfe *et al.*, 1989), divided by 2 to account for the parallel evolution of seed and CSL from their most recent common ancestor; the results were then scaled to millions of years and plotted to study their distribution (Figure **8**). The divergence times estimated for *N. benthamiana* show that the last-diverging sections were *Sylvestres, Petunioides* and *Noctiflorae*; however, as for the phylogenetic distance distributions, their divergence times distributions were largely overlapping. Based on the divergence time distributions obtained from trees having either one of the two parents as CSL, we would estimate the hybridization event to have taken place 4-5 million years ago (peak: 4.9 MyA). In *N. tabacum* we observed a clear separation between the distributions of the presumable two parental sections and all other sections leading to an estimation for the hybridization event having occurred 0-0.5 million years ago (peak: 0.4 MyA) (Table **3**).

Discussion

The progenitors of N. benthamiana

By means of a phylogenetic analysis, we have set out to identify extant species which are the closest living relatives of the parental progenitors of *N. benthamiana*. We studied a phylome constructed with gene sets from *N. benthamiana* and six diploid tobacco species. In a majority of cases, a homolog from section *Noctiflorae* represented the CSL, followed by section *Sylvestres*

(Figure **3a**). The strength of the hybridization signal from these candidate parental progenitors was clearly visible in the cluster network (Figure **4**). Previously, the maternal parent of *N. benthamiana* has been attributed either to section *Noctiflorae* or *Petunioides*, or to one of the two having introgressed DNA from the other (Clarkson *et al.*, 2004; Kelly *et al.*, 2012). In our analysis, we provide evidence that a *Noctiflorae* species is the most probable maternal subgenome donor of *N. benthamiana*. However, we do note the presence of a plateau in the distribution of phylogenetic distances obtained from section *Noctiflorae*. The plateau encompasses the peaks corresponding to sections *Sylvestres* and *Petunioides*. We speculate that the maternal progenitor might have introgressed some DNA from an ancestor of section *Petunioides* before hybridizing with an ancestor of section *Sylvestres*, in line with previous findings (Kelly *et al.*, 2012). We noted a disproportion in the two parental CSL counts for *N. benthamiana* which could be attributed to biased fractionation of the paternally inherited subgenome (Leitch *et al.*, 2008; Petit *et al.*, 2010; Renny-Byfield *et al.*, 2011). This is in line with our observation on subgenome intermixing, where we found that *N. sylvestris* reads cover only 10-20% of the non-repetitive positions in Nb-1 scaffolds. In the comparatively young allopolyploid *N. tabacum*, we did not observe such a trend: the sets of genes attributed as being derived from *Sylvestres* or *Tomentosae,* respectively*,* are similar in number (Figure **3b**) and the two subgenomes can be clearly separated as inferred from mapping results (Figure **6c** and **6d**). The phylogenetic distances of *N. benthamiana* orthologs to all the other tested species were in a narrow range (Figure **3c**). Even though the CSL counts showed a prevalence of section *Noctiflorae* over *Petunioides*, the close evolutionary distance between these two candidate parental species required a second line of evidence in order to strengthen our claim. When looking at the results obtained from mapping of mRNA-Seq reads onto the Nb-1 annotated coding regions, we could clearly choose section *Noctiflorae* over *Petunioides* (Figure **5**). Taken together, our results suggest that a species from the section *Noctiflorae* was very likely the maternal subgenome donor in the hybridization event that resulted in section *Suaveolentes*.

Subgenome separation

It has been shown that a 'genomic shock' (McClintock, 1984) induced by hybridization often leads to genomic rearrangements (Bashir *et al.*, 2018) paired with high retrotransposon activity (Petit *et al.*, 2010), and that, in *Nicotiana*, these dynamics correlate with hybrid age, beginning shortly after hybridization (Lim *et al.*, 2007). It is therefore expected that a hybrid as old as *N. benthamiana* would show extensive subgenomic intermixing, similar to hybrids of section *Repandae* (Lim *et al.*, 2007). Traditional methods based on read mapping were successful for subgenome separation of *N. tabacum* but failed for *N. benthamiana* (Figure **6**). In young hybrids, subgenomes are more easily distinguishable due to short divergence time. We assume that intermixing of subgenomes progresses with time, and in the older hybrid *N. benthamiana* reached a level at which subgenome discrimination is far less obvious, also because both the subgenomes and the extant relatives of their parental species have been diverging for all this time. To overcome this obstacle, we used the

parental assignment of the genes in more than 40,000 trees in the *N. benthamiana* phylome, based on which we inferred parental origin for 85% of the Nb-1 assembly. The observed disproportion in favor of maternally inherited DNA (50.3% vs 34.4%) may be a result of biased fractionation of the genome in its long-term diploidization process (Clarkson *et al.*, 2005). We concluded that subgenome separation based on large phylogenetic data collections can overcome the difficulties encountered by read mapping methods.

Dating of the interspecific hybridization event

For dating the *Suaveolentes* hybridization event we counted substitutions at neutrally evolving positions between orthologs in each of the trees used for counting CSL occurrences. We focused on fourfold degenerate sites (Crick, 1968). We first looked at transversions (Tv) and transitions (Ti) at these sites. It is generally accepted that Tv are less frequent than Ti; the accumulation of transversions at fourfold degenerate sites (4DTv) is therefore slow and can be used as a footprint of time. Moreover, it is expected that 4DTv ratios computed from extant species related to the parental progenitors are similar, since both progenitors have been separated from the hybrid for the same time. As with phylogenetic distances, our *N. benthamiana* results could not clearly separate the parental progenitors from the others (Figure **7**). We note that sections *Trigonophyllae* and *Tomentosae* have a slightly larger 4DTv ratio when compared to the other four sections. Interestingly, section *Paniculatae* (represented by *N. cordifolia*) shared the same 4DTv profile with *Noctiflorae, Petunioides* and *Sylvestres*. However, this is likely to be an artifact, as *N. cordifolia* is endemic to islands off the Chilean coast (Clarkson *et al.*, 2017). In fact, despite the similar 4DTv profile, its phylogenetic distance distribution (Figure **3c**) indicates larger evolutionary distance. We speculate that sections *Noctiflorae, Petunioides* and *Sylvestres* diverged at about the same time. In the case of *N. tabacum,* its young age did not allow for many transversions to accumulate in the genome, therefore the peaks of the distributions regarding the parental species are close to zero. When looking at the divergence times as computed from the *N. benthamiana* phylome, it is also complicated to distinguish between parental and non-parental sections (Figure **8a**). We clearly identified a separate peak for section *Sylvestres*, but the peaks of sections *Noctiflorae* and *Petunioides* were close. From the other analyses in our work we know that the maternal progenitor likely was a species from section *Noctiflorae.* However, its peak and the one of section *Sylvestres* do not coincide. It can be envisaged that the genomes of *N. sylvestris* and *N. noctiflora* are only modern descendants of the species that generated the original amphidiploid hybrid. Hence, their genomes might be different from their ancestral counterparts, or could have mutated at different rates. Our results place the hybridization events leading to *N. benthamiana* at 4-5 MyA and to *N. tabacum* at 0-0.5 MyA (Figure **8**). The arithmetic means computed on the divergence times between each taxon and the hybrids (Table **3**) were in concordance with the recent literature (Clarkson *et al.*, 2017), which suggests an age of 6 MyA for section *Suaveolentes*, with maternal and paternal sections returning different estimations (6.4 MyA and 5.5 MyA, respectively). They

also show an age of 0.4 MyA for *N. tabacum,* which fits our findings. Taking these results into account, we conclude that *N. benthamiana* likely originated from a hybridization event which occurred around 5 MyA.

Methods

Gene sets and transcriptome assemblies

We selected one *Nicotiana* species as representative for each of the eight sections (Figure **1**, Table **1**). We obtained gene sets for *N. attenuata* and *N. obtusifolia* (Xu *et al.*, 2017), for *N. sylvestris* and *N. tomentosiformis* (Sierro *et al.*, 2013), *N. benthamiana* (Schiavinato et al., 2019), and *N. tabacum* (Edwards at al., 2017) from public resources (http://nadh.ice.mpg.de/NaDH/download/; http://bioinformatics.boku.ac.at/NicBenth/Download/; https://solgenomics.net/organism/Nicotiana_tabacum/genome). For *N. cordifolia* and *N. noctiflora* we assembled gene sets from publicly available transcriptome sequencing data. Illumina mRNA-Seq data from leaves (i.e. high-throughput sequencing data generated on polyA+ selected transcripts after conversion to cDNA) were downloaded from the NCBI short-read archive (SRA) (Leinonen *et al.*, 2011) for *N. cordifolia* and *N. noctiflora* (Table **1**). We quality-trimmed the reads with Trimmomatic v0.35 (Bolger *et al.*, 2014) (SLIDINGWINDOW:2:20 MINLEN:50) and assembled them using Trinity v2.4.0 (Grabherr *et al.*, 2011) (--no_normalize_reads --min_contig_length 200 --KMER_SIZE 25 --min_glue 2). A single transcript sequence, its coding sequence, and its protein translation were obtained for each gene with Trans-decoder v3.0.1 (Haas *et al.*, 2013), using the --single_best_orf option. Sequences translating a protein shorter than 100 amino acids were removed. The protein sequences of the assembled gene sets were used for validation with BUSCO v2.0 (Simão *et al.*, 2015) and the embryophyta odb9 dataset (-e 0.001 -m prot -f --long) downloaded from https://busco.ezlab.org/datasets/. For each of the eight gene sets, we represented each gene with its longest isoform only.

Phylome reconstruction

We generated a comprehensive collection of phylogenetic trees (“phylome”) for *N. benthamiana* and another one for *N. tabacum* using the phylomeDB pipeline (Huerta-Cepas *et al.*, 2014). Based on both coding sequences (nt) and protein sequences (aa), the pipeline performs homology searches, multiple sequence alignments (MSA) and generates a maximum likelihood (ML) tree for each gene of the seed species (Huerta-Cepas *et al.*, 2007; Huerta-Cepas *et al.*, 2014). A protein MSA is generated and quality-trimmed; the trimmed protein MSA is then converted to a nucleotide MSA using the nucleotide coding sequences of each gene so that silent mutations can be evaluated. We uploaded the protein sequences of *N. benthamiana* (Schiavinato *et al.*, 2019), *N. tabacum* (Edwards *et al.*, 2017)*, N. attenuata, N. obtusifolia* (Xu *et al.*, 2017)*, N. sylvestris,* and *N. tomentosiformis* (Sierro *et al.*, 2013) to the phylomeDB database, using the corresponding NCBI

taxonomy IDs (Table **1**, excluding *N. cordifolia* and *N. noctiflora*). Sequences yielding identical md5sums were collapsed upon upload. We specified either *N. benthamiana* or *N. tabacum* as the seed species to generate their corresponding phylomes. For *N. benthamiana* we first used the four published diploid gene sets (see above) to build a backbone phylome and then added the sequences of the two transcriptomes assembled by us, to reduce the risks of generating wrong topologies (Figure **2**). Then, we generated a blast database with the protein sequences of *N. attenuata, N. benthamiana, N. obtusifolia, N. sylvestris*, and *N. tomentosiformis*, and searched in this database for homologs to the newly assembled sequences of *N. cordifolia* and *N. noctiflora* using blastp v.2.2.30 (Camacho *et al.*, 2009) with thresholds for E-Value (1E-03), alignment overlap (0.3) and sequence identity (30%). The coding sequence of each gene that found a homologue within these thresholds was incorporated in the corresponding gene tree using MAFFT v6.861b (Katoh and Standley, 2013). The *N. tabacum* phylome, in contrast, was calculated by including both the published and the assembled sequences at the same time.

The output of the PhylomeDB pipeline was a collection of ML trees and their underlying nucleotide MSAs. Trees and alignments of the *N. benthamiana* phylome were stored under the entry "phylome 817" (http://phylomedb.org/phylome_817) in the PhylomeDB database, whereas those of the *N. tabacum* phylome were stored under the entry "phylome 251" (http://phylomedb.org/phylome_251), and can be downloaded or browsed.

Phylome analysis

Using custom python scripts based on ETE3 (Huerta-Cepas *et al.*, 2016) sequences that contained > 1% of undetermined bases ('N') were removed from the alignments underlying the gene trees to avoid similarity artifacts due to missing information; if the number of species in a tree fell below five in this step, we discarded the tree. From each phylome we then extracted the subset of trees that had: 1) at least five species represented, 2) at most three species in the sister group of the seed gene, 3) at most 0.1 phylogenetic distance (substitutions per site) between the seed gene and the closest sister leaf (CSL), and 4) at least a support of 0.9 at the most recent common ancestor (MRCA) between the seed protein and the CSL. We extracted the CSL in each of these trees, and its phylogenetic distance (Figure **9a**: CSL distance = $d_1 + d_2$; Figure **9b**: $d_4 + d_6$). We computed fourfold degenerate transversion (4DTv) ratios between seed genes and CSLs using transitions and transversions at fourfold degenerate sites extracted from MSAs. We retained only ratios computed from at least five substitutions in trees where the seed-CSL distance did not exceed 0.5.

To date the hybridization event we divided the total number of substitutions at fourfold degenerate sites (regardless of the nature of the substitution) by the total number of such sites. The value was converted to million years by dividing it by a substitution rate of 5e-09 per position per generation time, which is an accepted value for nuclear plant genes (Wolfe *et al.*, 1989).

Cluster network

From the trees used to count CSLs we selected trees containing all the taxa and trees where the CSL belonged to one of the sections *Noctiflorae, Sylvestres* or *Trigonophyllae* in case of the *N. benthamiana* phylome, and to one of the sections *Sylvestres, Trigonophyllae* or *Tomentosae* in case of the *N. tabacum* phylome, respectively. In each tree we retained only the phylogenetically closest homolog for each taxon. From each pool of trees we extracted 100 random trees and generated a cluster network using Dendroscope 3 (Huson and Scornavacca, 2012). We iterated this procedure ten times, generating ten networks. We rooted the *N. benthamiana* phylome trees with section *Tomentosae*, known to be the sister taxon of the whole *Nicotiana* genus (Knapp *et al.*, 2004, p.2004). The *N. tabacum* phylome trees were rooted using section *Paniculatae*, since a species affiliated with *Tomentosae* is the paternal genome donor of the smoking tobacco. This was done merely for practical reasons, as we only generated unrooted networks. In each network we raised the cluster consensus threshold as long as at least one hybridization branch could be seen in the network. We exported each network as a circular cladogram.

Processing and mapping of transcriptomic data

From SRA we downloaded the mRNA-Seq datasets derived from *Nicotiana* species *N. attenuata, N. benthamiana, N. cordifolia, N. noctiflora, N. obtusifolia, N. sylvestris,* and *N. tomentosiformis*, respectively (Table **1**, "SRA code"). Reads were quality-trimmed with Trimmomatic v0.35 (ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 SLIDINGWINDOW:2:20 MINLEN:50). From each file, we extracted eight million read-pairs and cropped them to a length of 50 nt. We mapped the reads against the *N. benthamiana* Nb-1 draft genome assembly (Bombarely *et al.*, 2012) with HISAT2 v2.1.0 (Kim *et al.*, 2015) (--score-min L,2,-0.4 --mp 5,5 --rdg 7,3 --rfg 7,3 -k 5 --no-mixed) and kept only primary alignments (samtools view -F 0x0100, version 1.3, (Li *et al.*, 2009)). We isolated covered positions within the annotated exonic Nb-1 regions only, and extracted the coverage per position using bedtools v2.27.1 (Quinlan and Hall, 2010) (bedtools genomecov -dz). We retained only positions with a minimum coverage of 4. We counted the base pairs that were uniquely covered by each of the tested species, or by pairs of them, using bedtools intersect.

Subgenome intermixing

Genomic paired-end read data were obtained from SRA for *N. glauca* (SRR6320052, SRR6320053, SRR6320054, SRR6320055, SRR6320056, SRR6320057), *N. sylvestris* (ERR274528), and *N. tomentosiformis* (ERR274540) (Sierro *et al.*, 2013; Khafizova *et al.*, 2018). Reads were quality-trimmed and cropped to a length of 100 nt with Trimmomatic (LEADING:15 TRAILING:15 SLIDINGWINDOW:4:20 AVGQUAL:20 CROP:100 MINLEN:50). The read length was chosen based on the sample with the shortest reads (i.e. those of *N. sylvestris* and *N. tomentosiformis*). The *N. glauca* reads were combined in a single pool. The same number of reads, i.e. 240 million read pairs, was extracted for each species. This number was chosen based on the

smallest sample (*N. glauca*). The *N. glauca* and *N. sylvestris* reads were mapped against the *N. benthamiana* Nb-1 assembly (Bombarely *et al.*, 2012), while the same *N. sylvestris* plus the *N. tomentosiformis* reads were mapped against the *N. tabacum* Nitab-v4.5 assembly in its scaffolded version (Edwards *et al.*, 2017). Reads were mapped with HISAT2 v2.1.0 (--no-softclip --no-spliced-alignment --mp 6,2) using different scoring functions in Nb-1 (--score-min L,0.0,-0.3) and Nitab-v4.5 (L,0.0,-0.12), which translate into five and two mismatches per 100 positions, respectively. We retained only primary alignments for each read (samtools view -F 0x0100), sorted the reads by genome coordinate (samtools sort), retained reads not overlapping transposable elements (TEs) (samtools view -L, using a bed file containing only positions outside of annotated TEs), and computed the coverage per position (samtools depth). We obtained coverage profiles over the non-repetitive regions and over annotated coding sequences (CDS), respectively, referring to the published annotations for Nb-1 (Schiavinato et al 2019) and Nitab-v4.5 (Edwards et al 2017). We computed the mean coverage and the covered fraction (i.e. fraction of positions covered by reads) in each scaffold from the coverage profiles. Given the high fragmentation of the Nitab-v4.5 assembly we excluded all scaffolds < 10 kbp (24.5% of the total size of the Nitab-v4.5 assembly, consisting of > 1 million fragments).

Subgenome separation

The parental leaves in each tree of each phylome were analysed with a custom python script based on ETE3 (Huerta-Cepas *et al.*, 2016). Given the prior detection of the candidate parents in each phylome (see “Phylome analysis” methods section), the script extracted all the leaves belonging to those parental sections (i.e. *Noctiflorae* and *Sylvestres* for *N. benthamiana*, *Sylvestres* and *Tomentosiformis* for *N. tabacum*). Those leaves were ranked by distance from the seed gene, and the one with the shortest phylogenetic distance was used to assign a parental origin to the gene. We note that the closest parental leaf can either be the CSL (Figure **9a**, distance: $d_1 + d_2$) or different from the CSL (Figure **9b**, distance: $d_3 + d_4 + d_5$). Genes without a parental assignment were termed “orphan genes”, and the parental origin was inferred from the parental assignment (if any) of the upstream and downstream gene in the genome. If the neighboring genes had the same parental assignment based on their phylome trees the assignment was transferred to the orphan gene; otherwise the gene was left as “orphan”. If only either one of the upstream or downstream genes was assigned to a parental section the script transferred the parental origin of this assignment. All the assigned genes within a scaffold were used to assign the scaffold to a parental origin: if at least 75% of its parentally-assigned genes showed the same origin the scaffold was assigned to that parent.

Software and hardware specifications

The code for this study was written in a series of Bash scripts, Python v2.7 scripts, and SoS Notebooks (Peng *et al.*, 2018); the Notebooks featured kernels for Python v3.6, R v3.4.0, and the

Bash shell. All plots were generated using ggplot2 (Wickham, 2016). The phylomes were generated on an HPC cluster, using a node consisting of 64 cores (2.70 GHz) and 500 GB of RAM. The generation of one phylome took approximately two weeks of continuous computation using all the available cores. The phylome analyses were performed on one node featuring 24 cores (2.40 GHz) and 500 GB of RAM.

Data availability

Trees and alignments for each gene of the *N. benthamiana* phylome are stored under the entry "phylome 817" in the PhylomeDB database (http://phylomedb.org/). Trees and alignments for each gene of the *N. tabacum* phylome are stored under the entry "phylome 251". Maternal, paternal and orphan genes are listed in Additional File 1. Parental assignments for scaffolds containing genes are listed in Additional File 2. NibSet-1 gene models can be downloaded from our data repository (http://bioinformatics.boku.ac.at/NicBenth/Download/).

Acknowledgments

This work was supported by the Austrian Science Fund FWF (Doctoral program BioToP, Project W1224). We thank Lukas Mach for discussions and comments on the manuscript. Part of the calculations were performed on the ANT compute cluster at the Centre for Genomic Regulation CRG) in Barcelona, Spain. TG, JCD, and HH designed experiments and supervised the study. MS analysed and interpreted the data. MMH provided help with phylome analysis. MS, HH, and JCD wrote the manuscript with contributions from TG and MMH. All authors approved the final version of the manuscript.

Author contribution

HH, JCD, and TG conceived the study and supervised analyses, MS designed experiments and analysed the data, MMH contributed to phylome analyses, MS, JCD and HH wrote the paper with input from TG and MMH, all authors approved the final manuscript.

Conflicts of interest

None declared.

Supporting materials legends

Additional File 1. Assigned parental origin of *N. benthamiana* genes.

Additional File 2. Assigned parental origin of *N. benthamiana* Nb-1 scaffolds.

References

Aoki, S. and Ito, M. (2000) Molecular Phylogeny of *Nicotiana* (*Solanaceae*) Based on the Nucleotide Sequence of the matK Gene. *Plant Biology*, **2**, 316–324.

Bally, J., Jung, H., Mortimer, C., Naim, F., Philips, J.G., Hellens, R., Bombarely, A., Goodin, M.M. and Waterhouse, P.M. (2018) The Rise and Rise of *Nicotiana benthamiana*: A Plant for All Reasons. *Annual Review of Phytopathology*, **56**, 405–426.

Bally, J., Nakasugi, K., Jia, F., et al. (2015) The extremophile *Nicotiana benthamiana* has traded viral defence for early vigour. *Nature Plants*, **1**, 15165.

Bashir, T., Chandra Mishra, R., Hasan, Md., Mohanta, T. and Bae, H. (2018) Effect of Hybridization on Somatic Mutations and Genomic Rearrangements in Plants. *International Journal of Molecular Sciences*, **19**, 3758.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A. and Martin, G.B. (2012) A Draft Genome Sequence of *Nicotiana benthamiana* to Enhance Molecular Plant-Microbe Biology Research. *Molecular Plant-Microbe Interactions*, **25**, 1523–1530.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Clarkson, J.J., Dodsworth, S. and Chase, M.W. (2017) Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in *Nicotiana* (*Solanaceae*). *Plant Systematics and Evolution*, **303**, 1001–1012.

Clarkson, J.J., Kelly, L.J., Leitch, A.R., Knapp, S. and Chase, M.W. (2010) Nuclear glutamine synthetase evolution in *Nicotiana*: Phylogenetics and the origins of allotetraploid and homoploid (diploid) hybrids. *Molecular Phylogenetics and Evolution*, **55**, 99–112.

Clarkson, J.J., Knapp, S., Garcia, V.F., Olmstead, R.G., Leitch, A.R. and Chase, M.W. (2004) Phylogenetic relationships in *Nicotiana* (*Solanaceae*) inferred from multiple plastid DNA regions. *Molecular Phylogenetics and Evolution*, **33**, 75–90.

Clarkson, J.J., Lim, K.Y., Kovarik, A., Chase, M.W., Knapp, S. and Leitch, A.R. (2005) Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (*Solanaceae*). *New Phytologist*, **168**, 241–252.

Crick, F.H.C. (1968) The origin of the genetic code. *Journal of Molecular Biology*, **38**, 367–379.

Edwards, K.D., Fernandez-Pozo, N., Drake-Stowe, K., et al. (2017) A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics*, **18**, 448.

Grabherr, M.G., Haas, B.J., Yassour, M., et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Haas, B.J., Papanicolaou, A., Yassour, M., et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*, **8**, 1494–1512.

Herpen, T.W.J.M. van, Cankar, K., Nogueira, M., Bosch, D., Bouwmeester, H.J. and Beekwilder, J. (2010) *Nicotiana benthamiana* as a Production Platform for Artemisinin Precursors. *PLoS ONE*, **5**, e14222.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M. and Gabaldón, T. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Research*, **42**, D897–D902.

Huerta-Cepas, J., Dopazo, H., Dopazo, J. and Gabaldón, T. (2007) The human phylome. *Genome Biol.*, **8**, R109.

Huerta-Cepas, J., Serra, F. and Bork, P. (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, **33**, 1635–1638.

Huson, D.H. and Rupp, R. (2008) Summarizing Multiple Gene Trees Using Cluster Networks. In K. A. Crandall and J. Lagergren, eds. *Algorithms in Bioinformatics*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 296–305. Available at: http://link.springer.com/10.1007/978-3-540-87361-7_25 [Accessed February 4, 2019].

Huson, D.H. and Scornavacca, C. (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology*, **61**, 1061–1067.

Katoh, K. and Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kelly, L.J., Leitch, A.R., Clarkson, J.J., Knapp, S. and Chase, M.W. (2012) Reconstructing the Complex Evolutionary Origin of Wild Allopolyploid Tobaccos (*Nicotiana* section *Suaveolentes*). *Evolution*, **67**, 80–94.

Khafizova, G., Dobrynin, P., Polev, D. and Matveeva, T. (2018) Nicotiana glauca whole-genome investigation for cT-DNA study. *BMC Research Notes*, **11**, 18.

Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Knapp, S., Chase, M.W. and Clarkson, J.J. (2004) Nomenclatural Changes and a New Sectional Classification in *Nicotiana* (*Solanaceae*). *Taxon*, **53**, 73.

Lagerkvist, U. (1978) “Two out of three”: an alternative method for codon reading. *Proceedings of the National Academy of Sciences of the United States of America*, **75**, 1759–1762.

Leitch, I.J., Hanson, L., Lim, K.Y., Kovarik, A., Chase, M.W., Clarkson, J.J. and Leitch, A.R. (2008) The Ups and Downs of Genome Size Evolution in Polyploid Species of *Nicotiana* (*Solanaceae*). *Annals of Botany*, **101**, 805–814.

Li, H., Handsaker, B., Wysoker, A., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lim, K.Y., Kovarik, A., Matyasek, R., Chase, M.W., Clarkson, J.J., Grandbastien, M.A. and Leitch, A.R. (2007) Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytologist*, **175**, 756–763.

Long, N., Ren, X., Xiang, Z., Wan, W. and Dong, Y. (2016) Sequencing and characterization of leaf transcriptomes of six diploid *Nicotiana* species. *Journal of Biological Research-Thessaloniki*, **23**, 6.

McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792–801.

Nei, M. and Rooney, A.P. (2005) Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics*, **39**, 121–152.

Peng, B., Wang, G., Ma, J., Leong, M.C., Wakefield, C., Melott, J., Chiu, Y., Du, D. and Weinstein, J.N. (2018) SoS Notebook: an interactive multi-language data analysis environment J. Kelso, ed. *Bioinformatics*, **34**, 3768–3770.

Petit, M., Guidat, C., Daniel, J., et al. (2010) Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytologist*, **186**, 135–147.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Renny-Byfield, S., Chester, M., Kovarik, A., et al. (2011) Next Generation Sequencing Reveals Genome Downsizing in Allotetraploid *Nicotiana tabacum*, Predominantly through the Elimination of Paternally Derived Repetitive DNAs. *Molecular Biology and Evolution*, **28**, 2843–2854.

Schiavinato, M., Strasser, R., Mach, L., Dohm, J.C. and Himmelbauer, H. (2019) Genome and transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔXT/FT. *BMC Genomics*, **20**, 594.

Sierro, N., Battey, J.N.D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S., Peitsch, M.C. and Ivanov, N.V. (2014) The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications*, **5**, 3833.

Sierro, N., Battey, J.N., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M.C. and Ivanov, N.V. (2013) Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biology*, **14**, R60.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Topal, M.D. and Fresco, J.R. (1976) Base pairing and fidelity in codon-anticodon interaction. *Nature*, **263**, 289–293.

Wang, X. and Bennetzen, J.L. (2015) Current status and prospects for the study of *Nicotiana* genomics, genetics, and nicotine biosynthesis genes. *Mol. Genet. Genomics*, **290**, 11–21.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Wolfe, K.H., Sharp, P.M. and Li, W.-H. (1989) Rates of synonymous substitution in plant nuclear genes. *Journal of Molecular Evolution*, **29**, 208–211.

Xu, S., Brockmöller, T., Navarro-Quezada, A., et al. (2017) Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc Natl Acad Sci USA*, **114**, 6133.

Figure Legends

Figure 1. Representation of the *Nicotiana* sections and hybrids. Tree branches are intended to show relationships but are not scaled to actual phylogenetic distance. The tree topology is based on a previous study **(Leitch** *et al.*, 2008). Black branches indicate the evolution of diploid taxa,

colored branches refer to the evolution of hybrid taxa.

Figure 2. Workflow used in the generation of the *N. benthamiana* phylome. The backbone phylome was generated from gene models annotated within the sequenced genomes of five species of tobacco, including *N. benthamiana* as the seed species (blue box). To the backbone phylome, two more species (*N. cordifolia*, *N. noctiflora*) were added based on genes obtained from transcriptome assembly. Sections are indicated below the species names. Software and procedures are included in grey boxes. Red boxes indicate phylogenetic tree collections.

Figure 3. a) Closest sister leaf (CSL) occurrences for the *N. benthamiana* phylome constructed with gene sets from different tobacco species and *N. benthamiana* genes as seed species. Colors represent different *Nicotiana* sections represented in the phylome; one species per section was included in the phylome. b) CSLs in the *N. tabacum* phylome. c) Distribution of phylogenetic distances expressed as substitutions per site, in a phylome based on a *N. benthamiana* seed sequence and the closest leaf for each other taxon. Colors distinguish the different *Nicotiana* sections comprised by the phylome; one species per section was included in the phylome. d) Distribution of phylogenetic distances in the *N. tabacum* phylome.

Figure 4. Circular cladograms showing hybridization scenarios between *Nicotiana* taxa. Branch length is not proportional to phylogenetic distance, i. e. only the network topology is shown. Grey lines show tree-like branches, red lines show hybridization branches. Names of hybrid sections (*Suaveolentes* and *Nicotiana*) are indicated in black, bold characters. Names of the parental sections are indicated in bold, colored characters. **a)** Cluster network for *N. benthamiana*; **b)** cluster network for *N. tabacum*.

Figure 5. a) Positions in annotated *N. benthamiana* exons with coverage from reads obtained by mRNA-Seq of different *Nicotiana* species (Table **1**). Colors represent different *Nicotiana* sections included in the analysis. "*Suaveolentes*" refers to the *N. benthamiana* reads aligning against the *N. benthamiana* genome, i. e. the highest number of positions that one expects to be covered. **b)** Positions covered in the parental pairs. Each line indicates a separate pair, composed of section *Sylvestres* plus another indicated section. Black color flags positions covered by both sections, dark red indicates positions that are unique to section *Sylvestres.* The remaining part is unique to the particular indicated section (coloring as in **7a**).

Figure 6. Coverage analysis in *N. benthamiana* **(a, b)** and *N. tabacum* **(c, d)** scaffolds. Each dot represents a scaffold, colored according to the fraction of its non-repetitive sequence that is covered by parental reads. The coverage on the y-axis represents the scaffold's mean coverage, computed on positions with at least two mapped reads. The horizontal dashed line is the expected

coverage (15.3x). Although some scaffolds showed a coverage higher than 80x, the plot has been limited to the 0-80x range to improve readability. **a)** Coverage in non-repetitive regions of the *N. benthamiana* Nb-1 draft genome assembly. **b)** Coverage in CDS regions of the Nb-1 assembly. **c)** Coverage in non-repetitive regions of the *N. tabacum* Nitab-v4.5 assembly. **d)** Coverage in CDS regions of the Nitab-v4.5 assembly.

Figure 7. Ratio of transversions (Tv) at fourfold degenerate (4D) sites computed between a seed sequence and its closest sister leaf (CSL) in each phylogenetic tree of phylomes comprising different species of tobacco. Colors represent *Nicotiana* sections. The ratio is obtained by dividing the number of transversions to the total number of substitutions. **a)** phylome with seed sequence from *N. benthamiana* or **b)** with seed sequence from *N. tabacum*.

Figure 8. Distribution of divergence times (millions of years), obtained from the phylome trees. Colors represent *Nicotiana* sections. **a)** divergence time between *N. benthamiana* and a species of the indicated sections; **b)** *N. tabacum* results. The black dashed lines indicate the latest estimates for the timing when hybridization took place in the generation of sections *Noctiflorae* (a) and *Sylvestres* (b) (Clarkson *et al.*, 2017).

Figure 9. Schematic gene tree showing the seed gene, the closest sister leaf (CSL), and genes belonging to parental sections of the hybrid. Each colored circle represents a different species. Leaves in the phylogenetic tree belonging to the candidate parents are represented in the two circles with black borders. Each *d* represents a branch length (i.e. phylogenetic distance) relevant in our analyses. **a)** The CSL belongs at the same time to one of the parental sections. **b)** CSL and the closest parental leaf are different from each other.

Tables and Table legends

				Transcripts			mRNA-Seq data
Species	**NCBI Taxid**	**Section**	**Ploidy level**	**Raw**	**Analysed**	**Removed**	**SRA code**
N. attenuata	49451	*Petunioides*	2	33,449	32,968	481	SRR1950890
N. benthamiana	4100	*Suaveolentes*	4	50,503	50,090	413	SRR7540371
N. cordifolia	140890	*Paniculatae*	2	22,251	22,242	9	SRR2106516
N. noctiflora	118707	*Noctiflorae*	2	22,940	22,928	12	SRR2106514
N. obtusifolia	200316	*Trigonophyllae*	2	28,147	27,794	353	SRR2912995
N. sylvestris	4096	*Sylvestres*	2	39,450	32,877	6,573	ERR274390
N. tabacum	4097	*Nicotiana*	4	69,500	69,323	177	-
N. tomentosiformis	4098	*Tomentosae*	2	35,770	31,121	4,649	SRR2106531

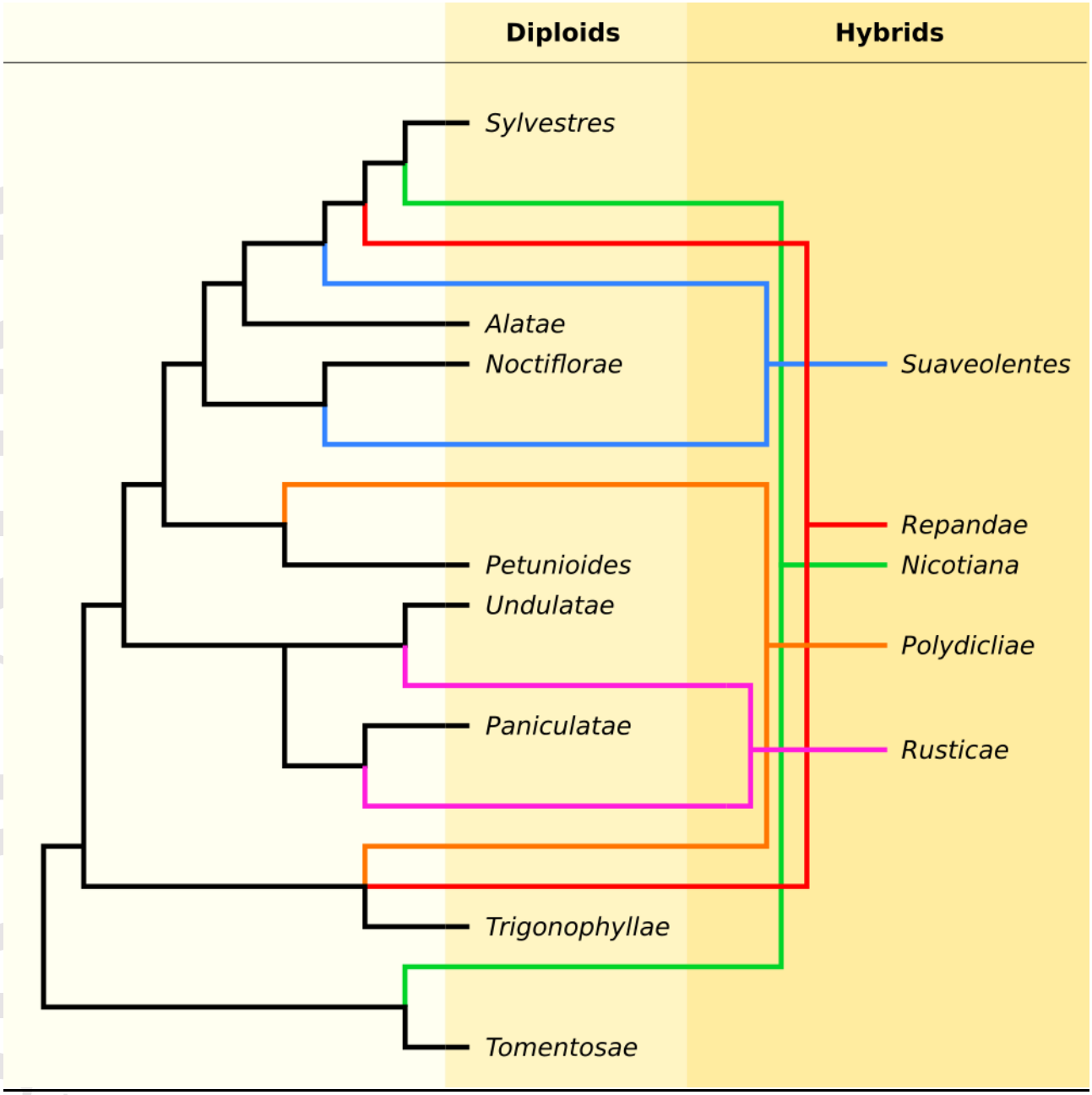
Table 1. Species and data representing the eight sections of the genus *Nicotiana* analysed in this study. Column “Analysed” indicates the number of genes uploaded into PhylomeDB for phylome construction. Identical sequences were merged (column “Removed”). SRA codes correspond to the raw mRNA-Seq data used for the transcriptomic coverage analysis and for gene set calculation (in case of *N. cordifolia*, *N. noctiflora*).

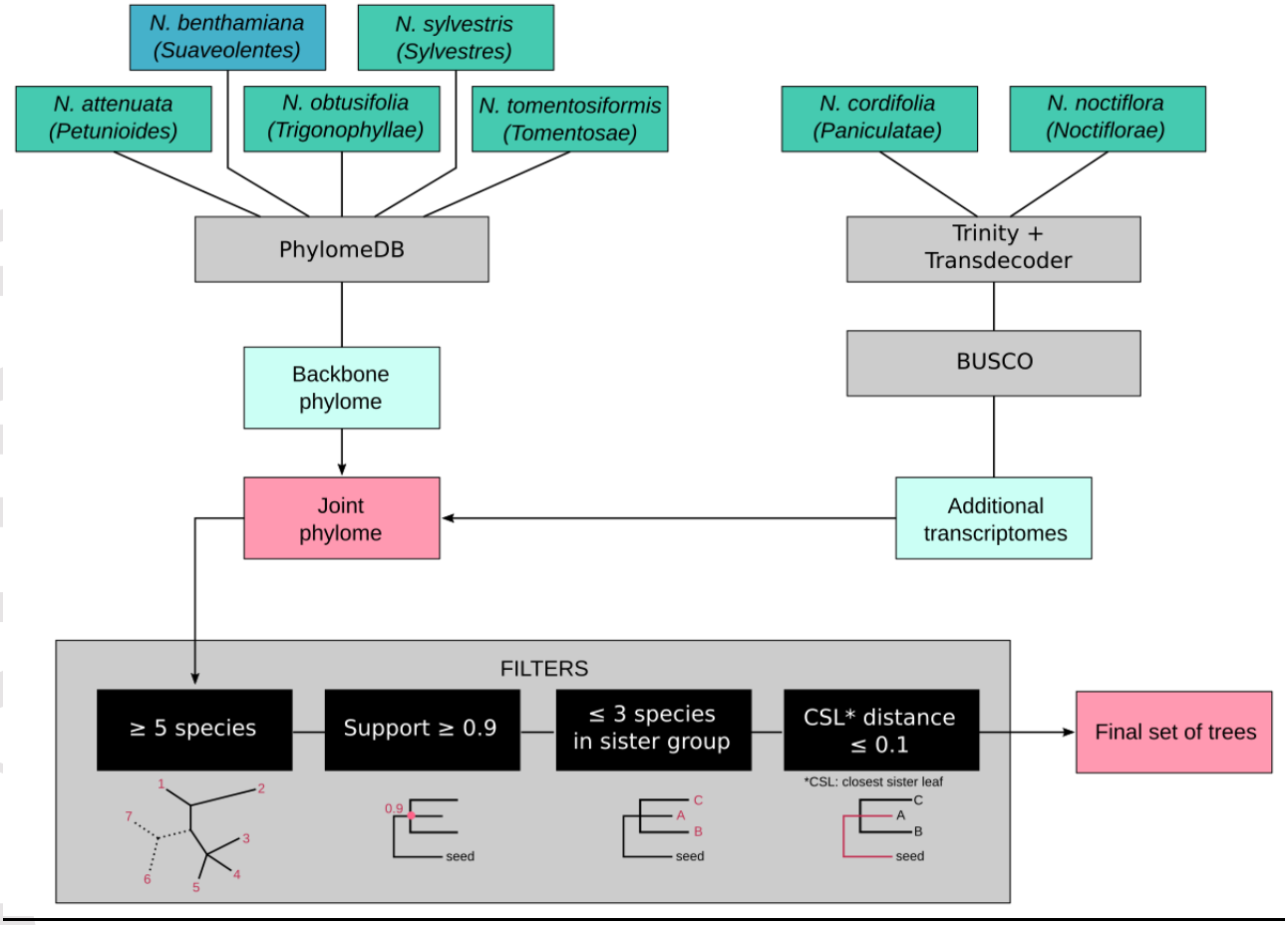
Assignment	**Number of scaffolds**	**Total length**	**Fraction (%)**	**Avg. scaffold length**	**Avg. number of genes**
Noctiflorae	3,972	1,419,752,456	50.3%	357,440.2	6.2
Sylvestres	3,259	970,554,883	34.4%	297,807.6	5.4
Orphan	3,132	432,579,414	15.3%	138,116.0	2.6
Total	10,363	2,822,886,753	100.0%	272,400.5	4.9

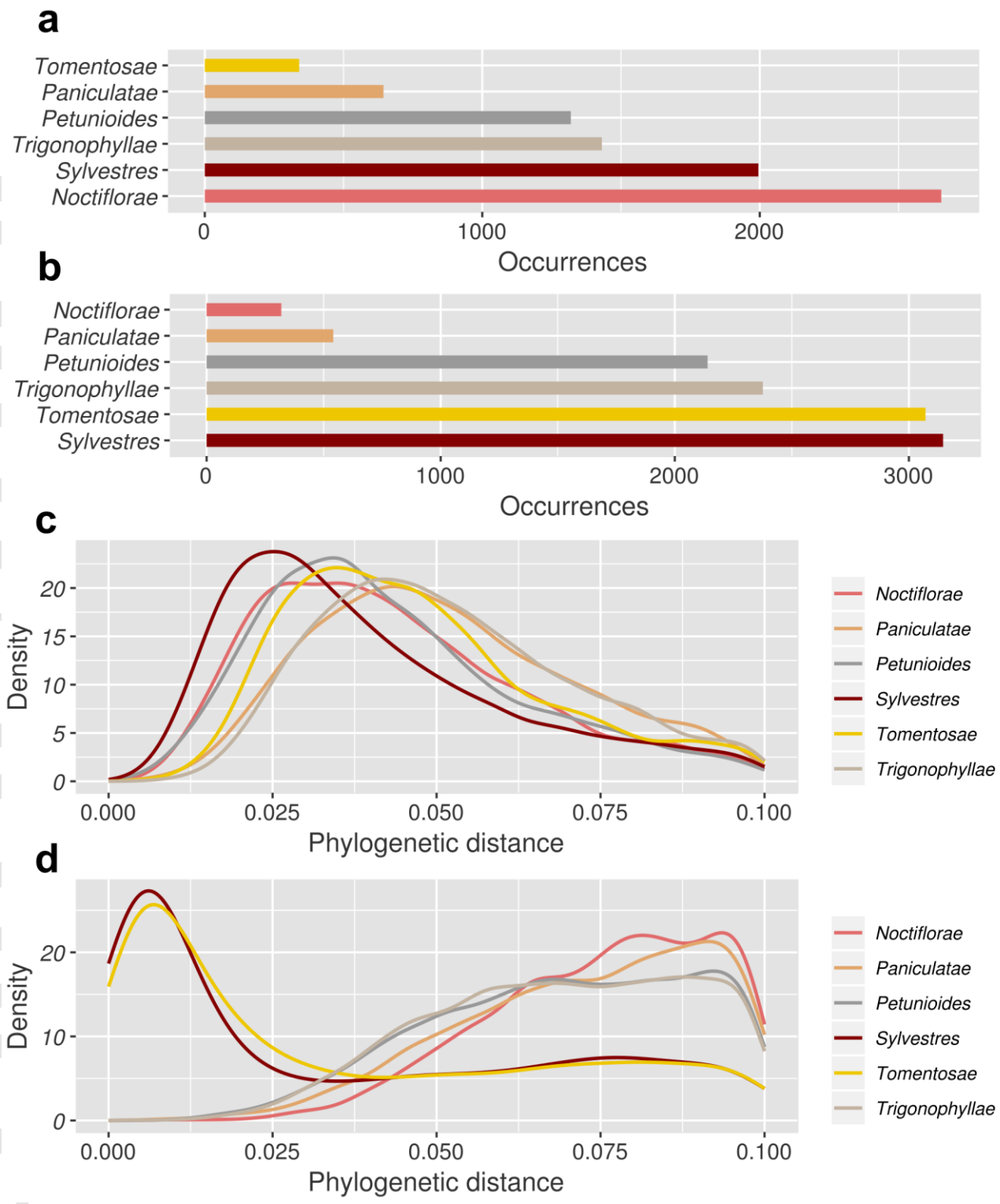
Table 2. Parental assignment of *N. benthamiana* scaffolds.

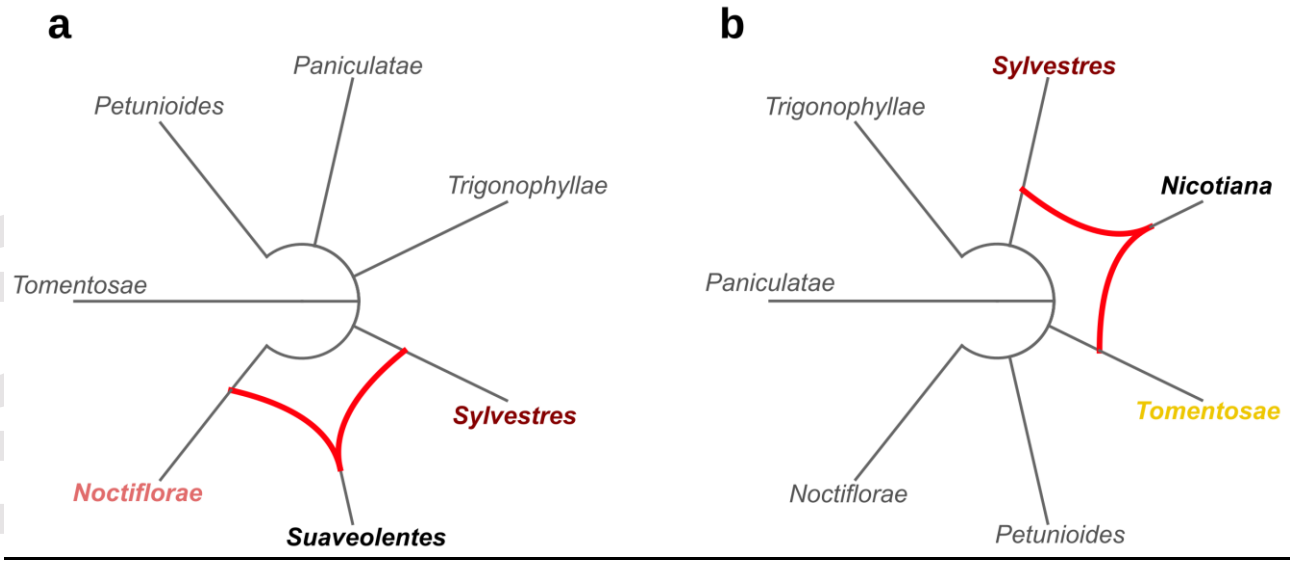
Section	***N. benthamiana* (peak, MyA)**	***N. tabacum* (peak, MyA)**
Noctiflorae	4.9	4.0
Paniculatae	6.1	3.0
Petunioides	4.7	3.1
Trigonophyllae	6.1	3.3
Tomentosae	5.8	0.4
Sylvestres	4.0	0.4

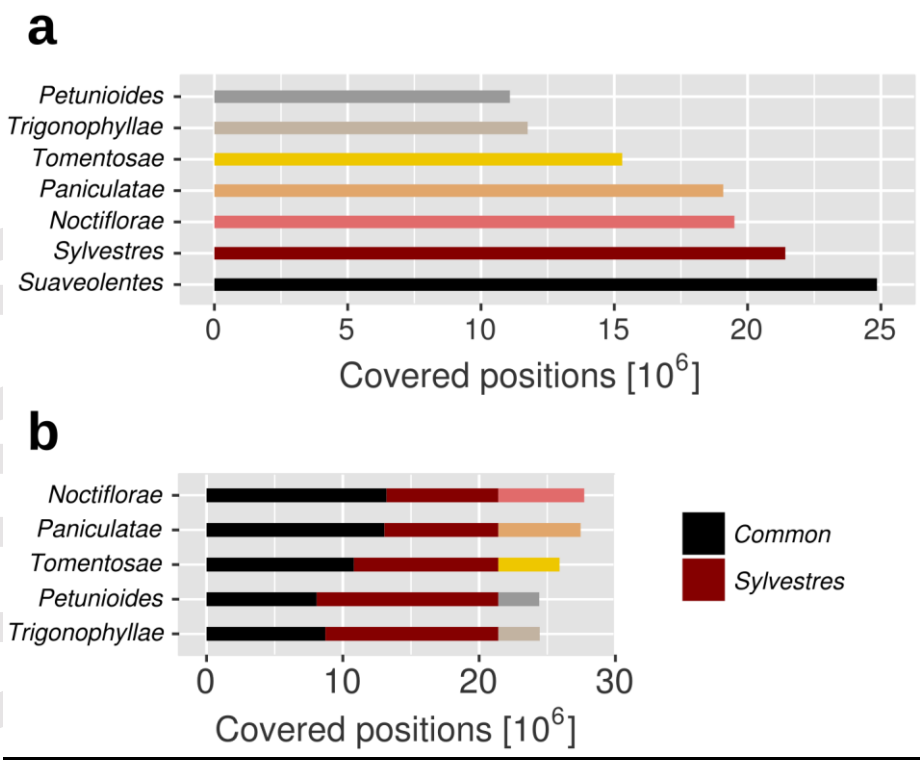
Table 3. Arithmetic means of the divergence time in million years obtained from each phylome by comparing the seed sequence against the closest homolog from each taxon in the data set.

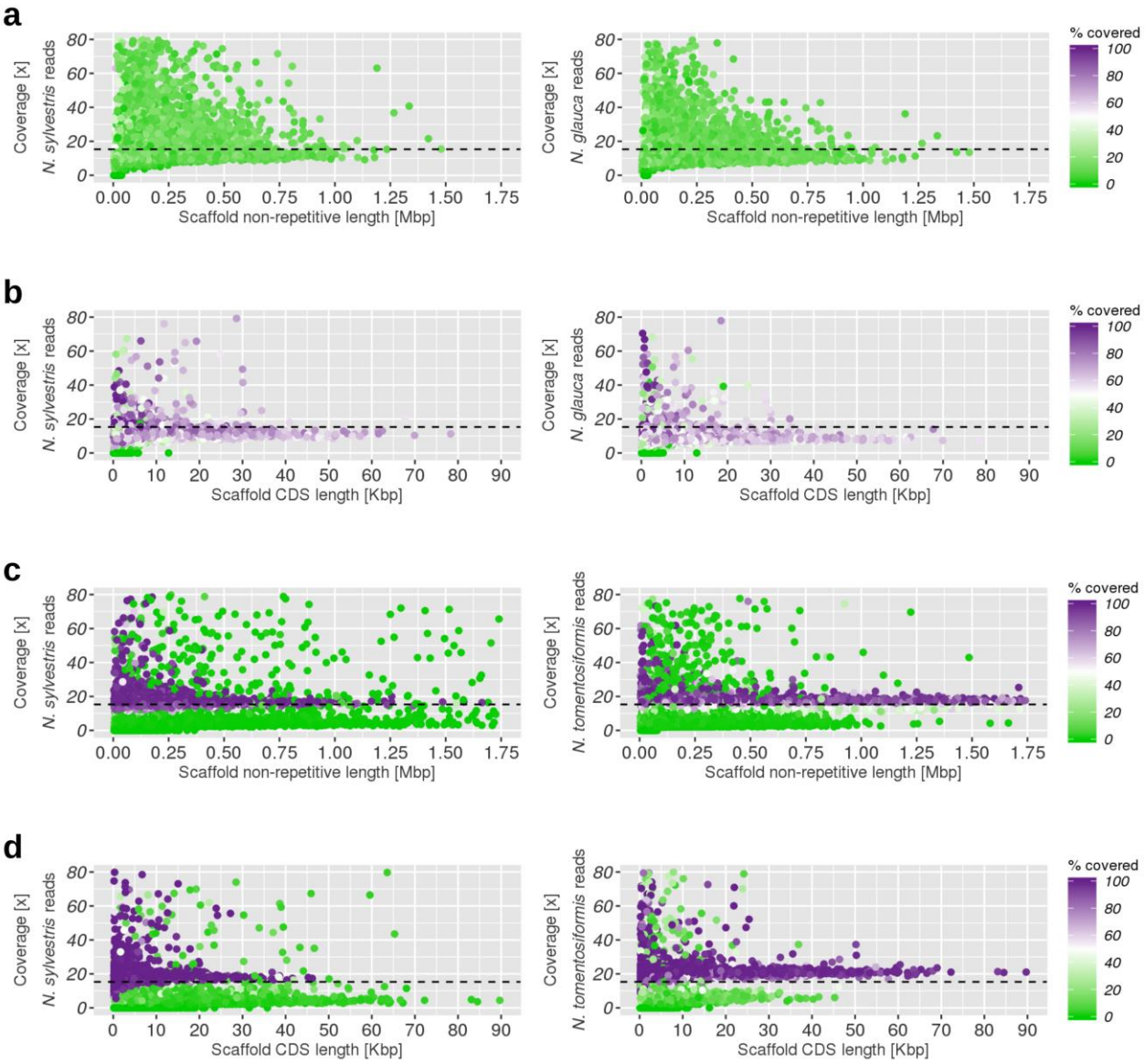


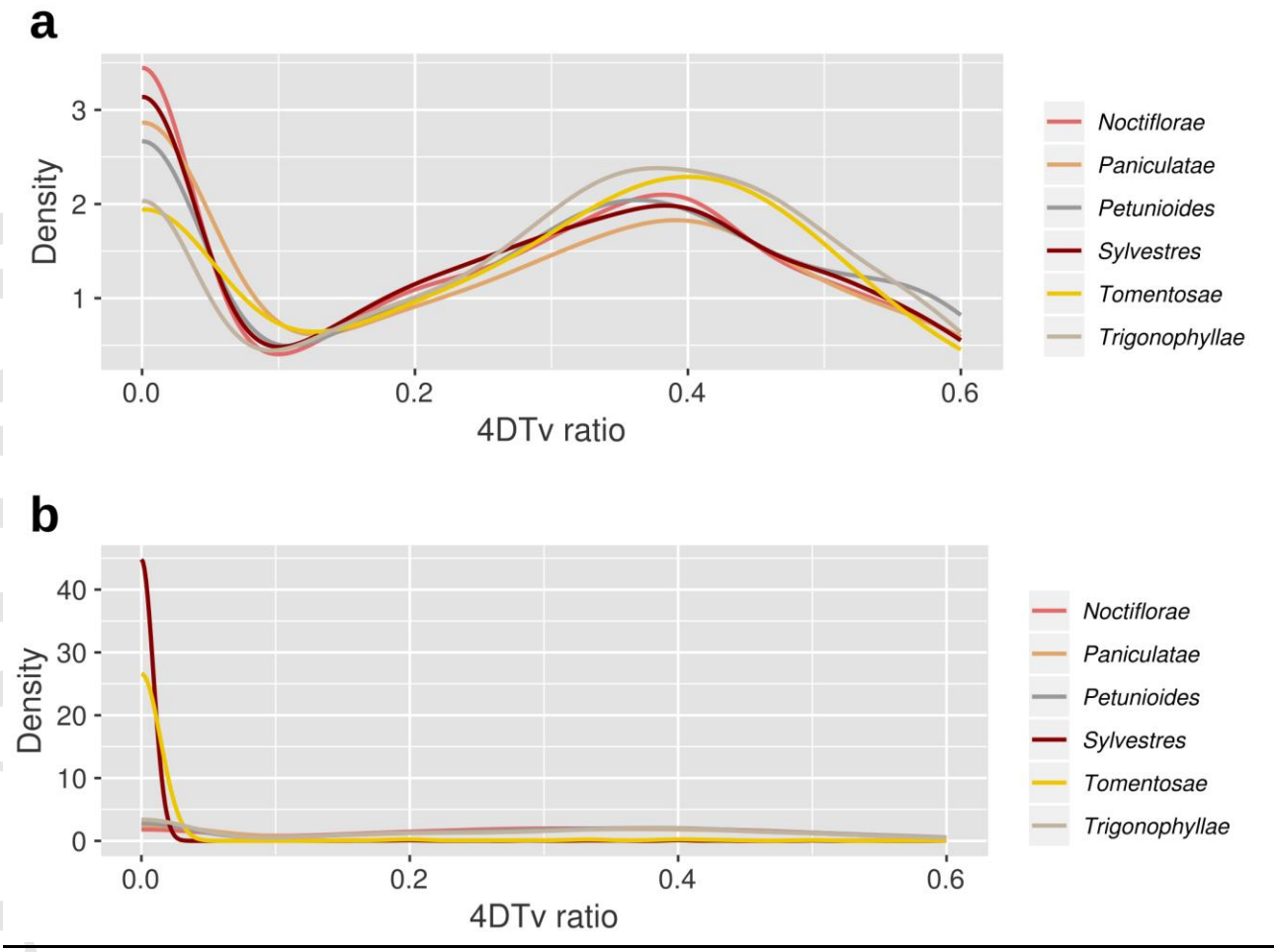


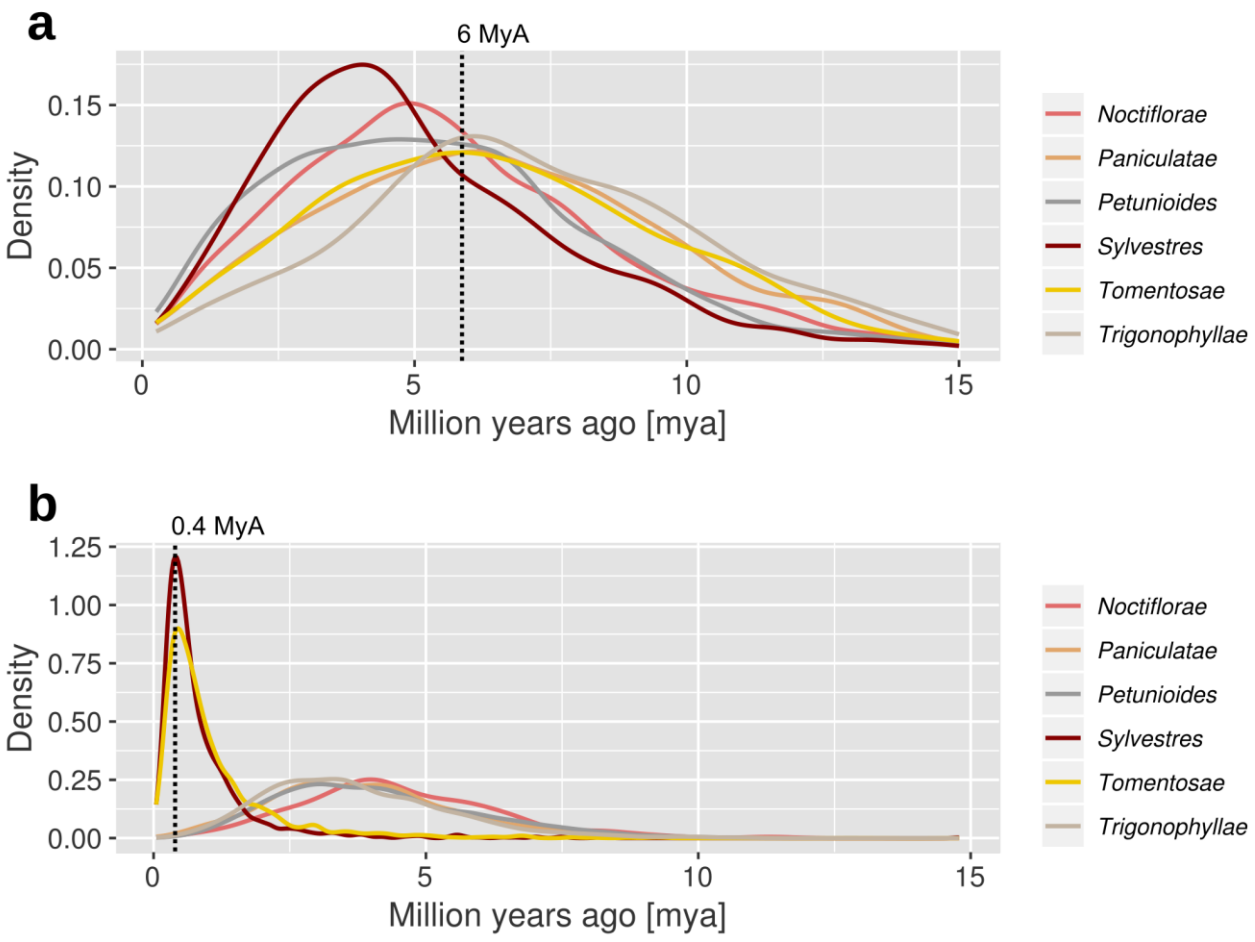


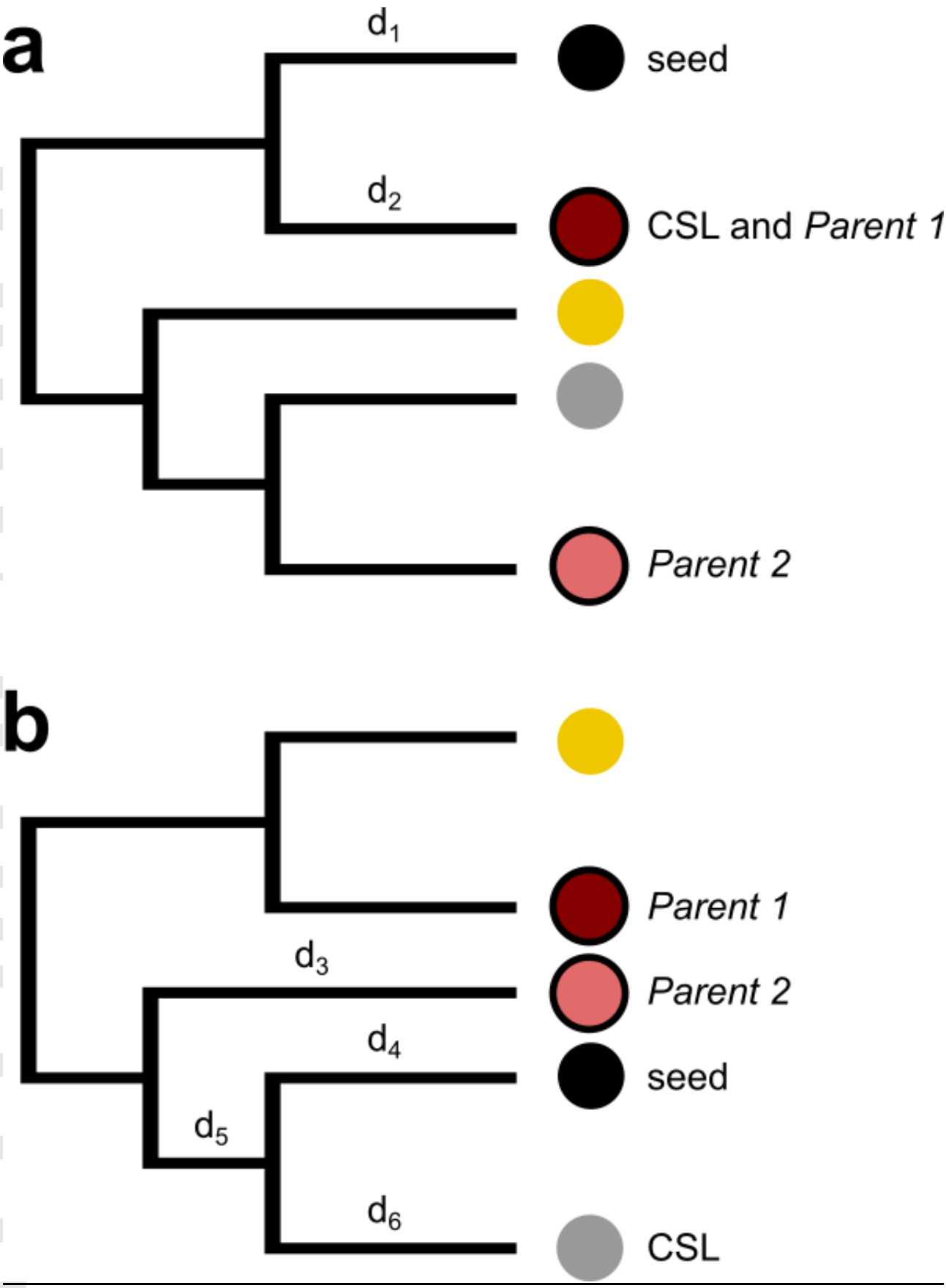












Chapter 4:

Conclusions

This work represents a comprehensive study of the genome of *N. benthamiana*. An in-depth gene prediction performed on the Nb-1 draft genome assembly (Bombarely et al., 2012) was produced, featuring 50,516 validated functionally annotated gene models. This gene prediction represented the groundwork of the entire thesis, and since it is publicly available, also a valuable resource for other researchers. The functional annotation, however, showed how much work is still to be done. Only 71% of the transcriptional isoforms were functionally annotated, while the remaining 29% did not satisfy our criteria or no homolog was found. This is the direct consequence of a lack of information that extends not only to *N. benthamiana,* but also to its closest relatives, which sensibly contributed to our annotation database. In the future, other studies will have to bridge this gap with reliable predictions and annotations, making use of the constantly increasing amount of data.

In this thesis I also analysed the ΔXT/FT research line (Strasser et al., 2008) and its relatedness with other research lines. Understanding the insertion position of the transgenes in ΔXT/FT showed a fusion between the ORF of one transgene and an endogenous gene, producing a chimeric transcript. Other copies of the disrupted gene were found in the genome, and no obvious differentially-expressed gene was observed that could harm recombinant protein production. However, these analyses must be performed in every such study, in order to clarify whether an inserted transgene is indeed fully functional, and the recombinant protein is the correct one. The relatedness of different research accessions was also addressed. As discussed in the introduction, the lack of documentation represented an issue in understanding the genome diversity between research lines of *N. benthamiana*. Here I showed that the variation density observed between accessions is compatible with them originating from one single source. The candidate accession is the one collected at the Australian Granites sites, already suspected to be the original specimen (Bally et al., 2018, 2015; Goodin et al., 2008).

Another important resource produced in this study has been the phylome containing the homology relationships between *N. benthamiana* genes and several diploid *Nicotiana* species. The value of this resource has been shown throughout the associated publication, where it was used for multiple purposes: to study the most likely parental progenitors; to date the hybridization event; to assign genes to their corresponding parental origin. Future studies will likely benefit in many ways from this publicly available resource.

In summary, this thesis represents a substantial advancement in the field of *N. benthamiana* genomics*,* and at the same time showed the power of high-throughput bioinformatic analyses when studying complex plant genomes.

Chapter 5:

Appendix

List of publications

Schiavinato, M., Strasser, R., Mach, L., Dohm, J.C. and Himmelbauer, H., Genome and transcriptome characterization of the glycoengineered *Nicotiana benthamiana* line ΔXT/FT. *BMC Genomics* 20, 594 (2019) doi:10.1186/s12864-019-5960-2

Schiavinato, M., Marcet-Houben, M., Dohm, J.C., Gabaldón, T. and Himmelbauer, H. (2019), Parental origin of the allotetraploid tobacco *Nicotiana benthamiana*. *Plant J.* Accepted Author Manuscript. doi:10.1111/tpj.14648

MATTEO SCHIAVINATO

PhD Candidate

DETAILS

matteo.schiavinato.90@gmail.com

SKILLS

- Genomics
- Python programming
- Plant Biology
- Bash programming
- R programming
- Linux operating systems

LANGUAGES

- Italian
- English
- German
- Spanish

HOBBIES

Musician, Sports lover, Basketball player, Martial Artist

EDUCATION

Primary and Secondary School, Montebelluna, TV, Italy
September 1996 — July 2009

Attended lectures on Energy Engineering, Università degli Studi di Padova, PD, Italy
October 2009 — September 2010

Bachelor of Science (B.Sc.) in Molecular Biology, Università degli Studi di Padova, PD, Italy
October 2010 — September 2013

Master of Science (M.Sc.) in Molecular Biology, Università degli Studi di Padova, PD, Italy
October 2013 — September 2015

Ph.D. candidate in the Biomolecular Technology of Proteins (BioToP) doctoral school, Universität für Bodenkultur (BOKU), Vienna, Austria
January 2016 — February 2020

INTERNSHIPS

Visiting Scientist at Centre for Genomic Regulation (CRG), Barcelona, Catalunya, Spain
June 2018 — December 2018

Acknowledgements

Firstly, I want to thank Professor Heinz Himmelbauer and Professor Juliane Dohm for believing in me and giving me this chance. Although I did not make it the first time, you offered me a second chance and I am grateful for that. During the last four years I learned many invaluable skills and, hopefully, I became independent in my research. I had the chance to work side by side with amazing people and made friendships that will hopefully last longer than this project. I would like to especially thank Alexandrina for enduring my rants about malfunctioning software and for making sure that a hot cup of coffee was always on the way. You are the sister I never had!

My second big “thanks” goes to my girlfriend Martina, for not letting me give up when I was over with it, not letting me become complacent when things went well, and letting me indulge in my love for music, sports and Shiba Inu breeds. I love you!

I am very thankful to Toni Gabaldón and his research group for hosting me in my stay-abroad in Barcelona, in 2018. I arrived with almost negligible knowledge on phylogenetic trees and I left with thousands of them. A special *merci* goes to Marina Marcet-Houben for being so patient, and to Ricardo, Marcello, Fran, Eugenio and Laura for all the things we did, and we might be better off not talking about in this very space.

During the last four years I owe most of the experiences I made to the BioToP doctoral programme. This thesis is the result of the nice and constructive environment that our adoptive PhD-mothers Christa Jakopitsch and Margareta Furtmüller created for the students. BioToP allowed me to meet those that are now among my best friends. Thank you Flávio, Madhu, Leander, Michael, Manuela, Christophe, Alejandro and all the others. Now roll me a charisma check, DC 15.

Finally, a very special thanks goes to my family: my mother for raising me the right way and convincing me that I could become a scientist one day, back when I was 16; my childhood friends who are always there for me when I come back, even if only for a day; my grandparents and grandaunt for always being so supportive; my uncles because Forza Toro. All of you, you might have not contributed to the scientific content of this thesis, but you surely made it happen in all the other ways!

Finally, thanks to myself. Because as a wise queen once said: if you can’t love yourself, how the h**l you gonna love somebody else?