Comparative *Nicotiana* genomics shed lights on the evolution of adaptive traits in plants

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. Nat.)

vorgelegt dem Rat der
Biologisch-Pharmazeutischen Fakultät der
Friedrich-Schiller-Universität Jena

von Dipl. Bioinf. Thomas Brockmöller
geb. am 13.07.1986 in Neu Kaliß
Gutachter:

1. Prof. Dr. Ian T. Baldwin (Max-Planck-Institut für Chemische Ökologie, Jena)
2. Prof. Dr. Günter Theißen (Friedrich-Schiller-Universität, Jena)
3. Prof. Dr. Karl Schmid (Universität Hohenheim, Stuttgart)

Beginn der Promotion: 17. Dezember 2013

Dissertation eingereicht am: 24. März 2017

Tag der öffentlichen Verteidigung: 30. Oktober 2017
Table of Contents

1 Introduction.........................................................................................................................7
  1.1 Adaptive traits in plants.................................................................................................8
  1.2 Mechanisms contributing to genome evolution and adaptive traits............................9
    1.2.1 Gene duplications........................................................................................................9
    1.2.2 Transposable elements...............................................................................................12
  1.3 Using phylogenomics to reveal gene and genome evolution.........................................14
  1.4 Nicotiana attenuata as model organism..........................................................................15

2 Aim of the thesis..................................................................................................................19

3 Overview of Manuscripts & Author’s Contribution.............................................................20
  3.1 Manuscript I......................................................................................................................20
  3.2 Manuscript II...................................................................................................................22
  3.3 Manuscript III..................................................................................................................23

4 Manuscript I..........................................................................................................................25

5 Manuscript II..........................................................................................................................127

6 Manuscript III..........................................................................................................................159

7 Discussion............................................................................................................................175
  7.1 The genome assembly of N. attenuata as basis for phylogenomic analyses.................175
  7.2 Gene and genome duplications provide genetic material for the evolution
    of novel traits.......................................................................................................................177
  7.3 Transposable elements promote rapid expression divergence.....................................180
  7.4 A combination of gene duplications and subsequent TE insertions directly
    contribute to the evolution of complex biosynthetic pathways.........................................183
  7.5 NaDH to study adaptive traits in Nicotiana.......................................................................185

8 References.............................................................................................................................188

9 Summary...............................................................................................................................199

10 Zusammenfassung................................................................................................................203

11 Eigenständigkeitserklärung...................................................................................................207

12 Acknowledgment..................................................................................................................209

13 Curriculum Vitae................................................................................................................211
1 Introduction

Adaptation to local habitats, a process that is facilitated by adaptive traits that protect all organisms from environmental stresses, is essential for their survival. Understanding the evolution of adaptive traits has long been a central topic in evolutionary biology. However, because most adaptive traits evolved long ago, elucidating their evolutionary history remains a big challenge, as it requires precise genomic information among closely related species, detailed knowledge on the genetic basis of the adaptive traits, and the availability of sophisticated phylogenomic tools.

In plants, structurally diverse specialized metabolites that serve complex functions for ecological adaptation to biotic and abiotic stresses are classical examples. During the last decades, several model systems have been established to study the biosynthetic machinery and the ecological functions of these metabolites. *Nicotiana attenuata*, a diploid annual species from the Solanaceae family, is one example. Using reverse genetic tools, the detailed molecular mechanisms and fitness effects of several specialized metabolites have been investigated both under laboratory conditions and under *N. attenuata’s* native habitat. Nicotine, the signature alkaloid of *Nicotiana* spp., has been shown to function as a defensive neurotoxin against attacking herbivores. The well-studied ecological functions of defensive metabolites such as this and the available genome sequences of four closely related species including *N. attenuata* together provide a great opportunity for investigating the evolutionary processes of adaptive traits in plants.

Gene duplication and transposable element (TE) insertions, the two key motors in genome evolution that are particularly common in plant genomes, can both affect the function of genes with adaptive consequences. While whole-genome and local gene duplications can provide the raw material for the evolution of novel traits, the mobility of TEs can broadly remodel gene expression, by redistributing transcription factor binding sites, shaping epigenetic marks, and/or providing target sequences for small regulatory RNAs. Hence, the combination of gene duplications and TE activity is thought to facilitate the evolution of novel adaptive traits. However, the details of this process, in particular its role in the evolution and assembly of complex plant metabolic pathways, remain unclear.
1 Introduction

1.1 Adaptive traits in plants

Due to their sessile lifestyle, plants are forced to deal with the threat of abiotic and biotic stresses. Abiotic stresses are environmental conditions that have negative effects on the fitness of the organism and can be introduced through limited nutrient or water content and high UV rate, among other examples. Biotic stresses are disadvantageous conditions introduced by interactions with other organisms sharing the same environment, e.g. by interactions with herbivores, pathogens and competitors.

Under different abiotic and biotic stresses, plants have evolved a large spectrum of physiological and metabolic responses, including short-term physiological changes (phenotypic plasticity) and fixed genetic changes that are associated with adaptive traits. Such adaptive metabolic or morphological responses may increase plants’ survival rate or fertility. Examples of well-studied traits are cold and heat resistance (Chinnusamy et al., 2007; Qu et al., 2013) as well as anti-herbivore defenses (Zavala, 2004; Boeckler et al., 2011). Particularly in the case of anti-herbivore defenses, plants have evolved a large range of highly specialized metabolites such as nicotine and tropane alkaloids (Detzel and Wink, 1993; Steppuhn et al., 2004) which can reduce the growth and survival rate of their natural enemies (War et al., 2012). The astounding diversity of defense compounds were thought to be a consequence of the ongoing arm race between plants and herbivores (Whittaker and Feeny, 1971; Berenbaum and Zangerl, 2008) that started more than 350 million years ago.

Anti-herbivore defenses can be classified as direct/indirect and constitutive/induced defenses (War et al., 2012). Classical examples for direct defenses are toxic substances or hairs, whereas volatiles to attract enemies of the herbivore represent indirect defenses (Schuman et al., 2012). In contrast to constitutive defenses that are present during the whole lifetime of the plant, induced defenses are explicitly activated after herbivore attack, a process which depends on the ability of plants to recognize insect feeding (Arimura et al., 2011; War et al., 2012)(Ludwig-Müller et al., 1997; Textor and Gershenzon, 2009). In manuscript II, the detailed mechanisms of this process and their evolution of the early defense signaling are presented.

The evolution of adaptive traits is highly connected with environmental stresses acting on the organism, due to the relationship between environment, phenotype and
1.1 Adaptive traits in plants

genotype. The phenotype of an organism is characterized by gene expression and regulation according to the current environmental situation. In this environment, the fitness of the organism is measured by the survival and propagation of this phenotype. Therefore, the environment creates a selection pressure on genetic and molecular traits that define the possible phenotypes. Due to this relationship between environment, phenotype and genotype, it is essential to analyze the involved genes and genomes for uncovering processes that guide the evolution of adaptive traits. Here, chemical ecology and ecophysiology can help to frame the analysis and findings in the correct biological context. Although many transcriptional and molecular responses to abiotic and biotic stresses have been analyzed in detail for several decades (Glawe et al., 2003; Hilker and Meiners, 2006), the evolutionary processes leading to these traits have not been revealed in the same depth, particularly not for complex traits involving multigenic biosynthetic pathways, as is the case for many anti-herbivore defenses. Already in 1970, Ohno (Ohno, 1970) was convinced that gene duplication may be an important mechanism for innovation and new functions, and since then many studies in molecular biology have illustrated the potential of gene duplications and transposable elements in the evolution in a wide range of genomic novelties (Cao et al., 2016).

1.2 Mechanisms contributing to genome evolution and adaptive traits

1.2.1 Gene duplications

Gene duplication is one of the most important mechanisms leading to genomic changes and therefore to the evolution of novel traits. Overall, gene duplications can be classified into three different types based on the mechanisms by which they were created: whole-genome duplication (WGD), tandem duplication (TD), transposon-mediated duplication, and segmental duplication of chromosomes and retro-duplication. But overall, WGD and TD account for the majority of gene duplications in plants (Panchy et al., 2016).

- Whole-genome duplication (WGD) is a process that replicates an entire genome, resulting in an increased ploidy level. Often, diploidization events, during which large number of duplicated gene were removed and chromosomes were
rearranged, are found after WGD. In angiosperms, polyploidization as well as diploidization events occurred multiple times (Bowers et al., 2003; Pfeil et al., 2005; Tomato and Consortium, 2012). Interestingly, many polyploidal plant species are able to survive under extreme environments and can adapt to a broad range of habitats. It is hypothesized that polyploid events provide evolutionary and ecological advantages for these organisms (Snyder and Cheson, 2000; Soltis and Soltis, 2000).

- **Tandem duplication** (TD) is process by which a gene is duplicated and inserted next to its ancestral copy in the genome. In contrast to a WGD, TD results in the duplication of a small number of genes, in particular genes occurring in clusters, by homologous recombination of paralogous sequences, called “unequal crossing-over”. Recent studies predicted that on average around 9% of the gene content in plant species occurred through TD (Rizzon et al., 2006; Panchy et al., 2016). However, genes evolved through TD are much more frequent than genes evolved through WGD and are significantly responsible for copy number and allelic variations within populations (Clark et al., 2007).

After duplication events, duplicated copies can be removed from the genome or can be retained, and evolve new functions. Duplicated copies are removed due to different means. On the one hand, genes gradually accumulate mutations, such as point mutations that can introduce premature stop codons that may disrupt their initial functions and can lead to non-functional copies. These genes will be removed or silenced over successive generations due to a decreased selection pressure for maintaining them, because high fitness costs are associated with keeping additional gene copies and genetic regions in the genome which may be introduced (i) by carrying and replicating duplicated sequences, (ii) by gene expression and production of RNAs and proteins, (iii) by expressed proteins that might be involved in energy-related reactions, and (iv) by duplications that lead to imbalances in gene dosage which may include improper gene regulations or unwanted molecular functions (Adler et al., 2014), known as the “dosage balance effect” (Hughes et al., 2007). Studies have estimated that the half-life of a duplicated copy is between 3 to 7 million years, and that within 50
million years, more than 90% of the duplicated genes are removed from a genome (Wagner, 2001). Although around 65% of the genes in the sequenced genomes of land plants originated by duplication (Panchy et al., 2016), their number has not increased constantly over time, and even species with several rounds of gene or genome duplication events consist of a similar number of genes (Lynch and Conery, 2001; Maere et al., 2005). Although the majority of duplicated copies will be removed, several copies are kept and fixed. Different theories have been postulated to describe this processes (Ohno, 1970; Hittinger and Carroll, 2007; Des Marais and Rausher, 2008).

- **Neofunctionalization**: One copy retains the ancestral function while the paralogous gene is free from constrains and can to evolve to new function. This is the classical positive selection model used to explain gene retention (Ohno, 1970).

- **Subfunctionalization**: This type defines a duplication that is characterized by an unequal degeneration of the ancestral function by asymmetrical mutations in the duplicated copies. Therefore, both copies together are needed to fulfill the ancestral function and will be kept in the genome.

- **Escape from adaptive conflict**: This type is similar to subfunctionalization. However, the novel functions already exist in a single-copy gene before duplication. The original gene fulfills several functions, but can not evolve to specialize in one of them to increase its efficiency without decreasing the others. After the duplication, each gene copy can specialize on a sub-function of the parental gene, which can result in an overall improvement for the organism.

- **Dosage balance effect**: This theory postulates that the selection pressure acts on both copies after duplication to keep them in the genome to maintain a dosage balance. This dosage balance is characterized for example by keeping all duplicated gene from a pathway to increasing their biosynthetic flux (Hughes et al., 2007).

Gene duplications play essential roles in the process of adaptation and evolution of novel traits in plants. Prominent examples include changes of floral structures
1 Introduction

(Guttesson and Reuber, 2004; Kramer et al., 2004; Dong et al., 2005; Mondragón-Palomino et al., 2009), improvements in disease resistance (Hofberger et al., 2014; Jia et al., 2015) and adaptation to various biotic and abiotic stresses (Edwards and Dixon, 2005). However, it is still not clear how gene duplications contribute to novel multigene pathways in plants.

1.2.2 Transposable elements

In most eukaryotes, the activity of transposable elements (TEs) is another important mechanism that shapes genomic structures and regulatory sequences. TEs are endogenous virus-like elements that are able to change their genomic location autonomously or with the help of the gene repertoire of the host. Often, TEs are called “jumping genes”. Historically, TEs were long considered as “junk” or “parasitic” DNA without function. But in recent years, emerging evidence shows that TEs are important components of genetic diversity that contribute the process of adaptation (Biémont, 2010; Kim et al., 2012).

TEs are classified in three main groups: Class 1, Class 2 and Helitrons. Class 1 transposons are retrotransposons that translocate in a “copy and paste”-like manner, during which the genomic sequence of TE is first transcribed to mRNA then back-transcribed to cDNA and inserted back into the genome. In contrast, Class 2 transposons are only translocated but not copied, since they are cut from the genome and re-inserted at a different location. The third group of transposable elements are Helitrons. It is assumed that they use a “rolling-circle” mechanism for transposition (Kapitonov and Jurka, 2007); however, this type is not further discussed during this dissertation. A TE can be classified as “autonomous" if it can move/copy by itself and does not depend on other TEs. In contrast, “non-autonomous" TEs do not have this ability, because they often lack genes which encode the required reverse transcriptase or transposase. Retrotransposons are the most common and the most abundant TEs in plants. They can be further divided into two families mainly based on the presence or absence of long terminal repeats (LTRs) that ranges from 100bp to over 5kb. The most common types of LTRs in plant genomes are copia-like and gypsy-like elements which are classified based on sequence similarity and the order of encoded gene products; LTRs contain cis-regulatory sequences and RNA polymerase II (Pol II) within each of the two flanking
regions of the elements (Mager and Stoye, 2015). Due to their size, LTRs have the largest effect on genome size. The other group of retrotransposons are non-LTRs which consist of two sub-types, long interspersed elements (LINEs) and short interspersed elements (SINEs). These elements can also be found in high copy numbers in plant genomes (Schmidt, 1999). In contrast to LINE that contain an internal Pol II promoter located in their 5’ UTR region and an anti-sense promoter (Speek, 2001), SINEs are nonautonomous retroelements and contain sequence motifs that are capable of recruiting Pol III (Richardson et al., 2015).

TEs are particularly predominant in plant genomes, and can occupy up to 85% of total genome size in some species (Lee and Kim, 2014). TEs can transpose at high frequencies ranging from $10^{-3}$ to $10^{-5}$ per element per generation, which is in several orders of magnitude higher than the nucleotide-base substitution rate ($10^{-8}$ to $10^{-9}$ per nucleotide per generation) (Biémont and Vieira, 2006; Saha et al., 2008). The inserted TEs not only provide raw genetic material for the evolution of new ORFs, they also contribute to rewiring of the gene regulatory networks. Three known mechanisms facilitate this process. First, the insertion of TEs at the promoter region of genes can introduce functional elements like promoters, enhancers and silencers that can lead to expression changes in adjacent genes (Elrouby and Bureau, 2012; Makarevitch et al., 2015). TEs contain functional binding sites of transcription factors (TFs) and recent studies suggest that spreading TF-binding sites within the genome might be one general mechanism for changes in regulatory networks and their evolution, in particular, by activating genes under conditions they were not expressed before (Sundaram et al., 2014). Second, due to the special copy mechanism of TEs, closely related regions, including genes, can be copied and duplicated. A process that can result in duplicated genes, new gene isoforms or splice sites – another major factor leading to altered gene expression or novel functions (Xiao et al., 2008). TEs can contain splice-sites themselves and therefore can introduce different splicing pattern of genes; a process that is called “exonization”. Although, the contribution to novel functional protein coding genes seems speculative (Piriyapongsa et al., 2007), findings indicate that TE insertions can, interestingly, lead to novel genetic features in the organism. However, their impact on complex adaptive traits is still not resolved in detail.
1 Introduction

Besides the positive effects on genome evolution, TEs can also disrupt the reading frame of genes or their regulatory elements by insertions into or near genes, respectively, which lead to loss of function in the corresponding genes. Furthermore, TEs can lead to genomic instability resulting in higher number of chromosomal rearrangements and DNA damage (Bennetzen, 2000). To prevent the negative effects introduced by their high transposition rates, many TEs are epigenetically silenced (Tanurdzic et al., 2008), inhibited by RNAi (Kasschau et al., 2007) or suppressed by transposon specific enzymes (Bucher et al., 2012). However, some TEs can be re-activated by epigenetic modifications in the corresponding genomic regions induced by abiotic and biotic stresses, such as mechanical damage and pathogen attack (Wessler, 1996; Grandbastien, 1998), as well as by polyploidy events (Alpert, 2006). The first stress-induced activation of TEs was found in tobacco, in which the retrotransposon TntI was found to be activated by stress treatment (Pouteau et al., 1991). It is proposed that the stress-induced activation of TEs might lead to the evolution of stress-responsive adaptive traits of influenced genes (Makarevitch et al., 2015). In summary, TE insertions can lead to genetic novelties that may contribute to the evolution of adaptive traits.

1.3 Using phylogenomics to reveal gene and genome evolution

Phylogenomics combines techniques from the field of phylogenetics and comparative genomics to infer evolutionary processes, and is one of the most promising approaches for understanding the evolution of genomes and adaptive traits (W.J. Murphy, 2008). This process is achieved by detecting unique and shared features among different species and analyzing the origin and evolution of genes. In my dissertation, I used and developed phylogenomic tools to: (i) predict gene functions, (ii) investigate gene family evolution, and (iii) analyze the evolution of pathways.

For phylogenomic analysis, the first step is to infer gene families and define homologous groups among genes that share the same evolutionary history. Computationally, gene homology inference and gene family identification are defined based on sequence similarity information – usually based on protein sequences, functional motifs or protein domains (Kuzniar et al., 2008). The identified gene families can then provide several pieces of useful information on the function and evolution of
1.3 Using phylogenomics to reveal gene and genome evolution

the studied organisms. First, gene family identification can provide functional inferences to unknown genes. This is because genes that share similar sequence motifs tend to have similar functions; the so-called “guilt by association principle” (Eisen et al., 1998). Second, copy number fluctuations of certain gene families among species provide information on species- or lineage-specific expansions/reductions of gene content in organisms, which can shed light on the species-specific evolution of novel functions and traits (Jin and Martin, 1999; Lehti-Shiu et al., 2016). To precisely estimate gene loss and gain in each lineage, the tree reconciliation approach is often used. In this approach, the process of gene gain and loss are inferred by mapping a phylogenetic gene tree to the corresponding species tree (Nakhleh, 2013). However, due to horizontal gene transfer among different species, the nonlinearity of the molecular clock, unequal gene-gain and gene-loss among species, and variations of evolutionary speed among genes, the evolution of genes often does not coincide with the evolution of species, which complicates the process of tree reconciliation. To handle this, current approaches use statistical methods to infer the most likely phylogenetic gene tree under a given species evolution by using mathematical models. Often models use a cost function to indicate the likelihood of gene losses, gene duplications, and horizontal gene transfer based on biological knowledge (Chen et al., 2000).

1.4 *Nicotiana attenuata* as model organism

The genus *Nicotiana*, a sister clade to *Solanum* within the Solanaceae, consists of 75 naturally occurring species mainly native to America and Australia (Chase et al., 2003). A schematic representation of the phylogenetic relationships is presented in Figure 1.
Figure 1: Phylogenetic relationship of Nicotiana species within the eudicots. Red stars and green stars indicate previously characterized whole-genome duplications and whole-genome triplications, respectively.

The diploid (2n with 24 chromosomes) wild tobacco, *N. attenuata* [Torr. ex S.Watson] is native to the Great Basin desert of the United States with populations across Utah, Nevada, Arizona, Oregon and California. This species has been used extensively as a model organism to study plant-environment interactions and plant gene function under real-world conditions during the last decades (Wells, 1959; Baldwin, 2001; Kessler et al., 2010; Schuman et al., 2012). Conditions in this desert environment, such as high UV radiation levels, long drought periods, and highly dynamic herbivore communities, constantly challenges the survival of this species. In response, *N. attenuata* has evolved sophisticated adaptive traits that allow this plants to adapt to an ecological niche characterized by post-fire environments defined by high concentrations of nitrogen in the soil and a low number of competitor plant species (Baldwin et al., 1987; Baldwin and Morse, 1994). As a consequence of this fire-chasing behavior, *N. attenuata* usually grows in near-monoculture populations, which makes them a preferred food source for a wide-range of herbivores, such as the lepidopteran larvae of *Manduca sexta* and several leafhoppers. These herbivores result in highly dynamic biotic stresses that challenge the fitness of these plants in nature. As a consequence, *N.*
attenuata evolved several defensive traits that protect them from these various attacking herbivores.

Nicotine, a pyridine alkaloid, represents a key anti-herbivore defensive traits that is common to most Nicotiana species. Nicotine is toxic to most herbivores and is the most abundant alkaloid in N. attenuata (Steppuhn et al., 2004). It binds to the acetylcholine receptors in the nervous systems of animals, thus acting as a defense compound against herbivores. Nicotine in the flowers of N. attenuata also influences florivory, nectar robbing and pollinator foraging behavior (Kessler et al., 2008). The biosynthesis of nicotine involves genes from both nicotinamide adenine dinucleotide (NAD) cofactor and polyamine metabolism pathways. Nine genes are known to be involved in the nicotine biosynthesis: arginine decarboxylase (ADC), ornithine decarboxylase (ODC), putrescine N-methyltransferase (PMT) and N-methylputrescine oxidase (MPO) which encode for proteins involved in forming the pyrrolidine ring (Hibi, 1994; Imanishi et al., 1998; Bortolotti et al., 2004; Xu, 2004; Heim et al., 2007; Katoh et al., 2007), while aspartate oxidase (AO), quinolinate synthase (QS) and quinolinic acid phosphoribosyltransferase (QPT) are forming the pyridine ring (Sinclair et al., 2000; Katoh et al., 2006). It is proposed that both ring formations are combined by the genes isoflavone reductase-like protein (A622) and berberine bridge enzyme-like enzymes (BBL) (DeBoer et al., 2009; Kajikawa et al., 2009; Kajikawa et al., 2011). In manuscript I, using a phylogenomic approach, we show how gene duplications and TE insertions influenced the evolution of this highly specialized multigenic pathway.

In addition to nicotine, phenolamides represent another class of defensive metabolites in N. attenuata, which includes a group of phenolic derivatives that originate from the main phenylpropanoid pathway after conjugation to polyamines or aryl monoamines (Bassard et al., 2010). These compounds can be found in many plant lineages and have been long assumed to only be involved in the development of flower buds. However, recent studies showed that phenolamides also have a defensive function and are accumulated after herbivore attack, during which their biosynthesis is activated by the key transcription factor MYB8 (Onkokesung et al., 2012; Gaquerel et al., 2014). In manuscript III, using MYB8 as an example, we show how combinations of phylogenomics and co-expression analysis can facilitate the identification of new
1 Introduction

regulatory pathways and how such pathways evolved. The tools and data generated during my dissertation are implemented in the \textit{Nicotiana attenuata} data hub (NaDH), which provides open access for all users to analyze the function and regulations of genes and metabolites in tobacco.

Based on the comprehensive characterization of ecologically-relevant traits and the phylogenetic relationship to other well-known model organisms, \textit{N. attenuata} is an important model organism for understanding the evolution of adaptive traits, and provides the base for analyses performed during this thesis.
2 Aim of the thesis

The aim of my thesis is to shed light on the role of gene duplications and TE activity in the evolution of adaptive traits. Although, the evolution of adaptive traits is a central topic in evolutionary biology, the details of this process, and in particular the role of gene duplications and TE activity in the evolution of plant metabolic complexity through the assembly of novel multigenic pathways, remain unclear. For this reason, I combined established methods as well as developed new approaches in the field of phylogenomics to address the following questions:

Manuscript I:

- How did the *Nicotiana attenuata* genome evolve, and which genomic features differentiate *Nicotiana* from other *Solanaceae* species?
- What are the consequences of TE insertion in the genomes of *Nicotiana* spp.?
- How did the nicotine biosynthetic pathway evolve?

Manuscript II:

- Which genes are involved in herbivore-induced early defense signaling network in *Nicotiana* spp.?
- How did herbivore-induced early defense signaling evolve in *Nicotiana* spp.?
- How did gene duplications and TE insertions contribute to the evolution of herbivore-induced early defense signaling in *Nicotiana* spp.?

Manuscript III:

- Can co-expression analysis result in the identification of genes involved in specialized metabolites?
- How can we combine phylogenomic data with gene co-expression analysis to reveal the evolution of gene networks?
3 Overview of Manuscripts & Author’s Contribution

3.1 Manuscript I

“Wild tobacco genomes reveal the evolution of nicotine biosynthesis”


Submitted to PNAS (January 2017)

In manuscript I, we sequenced, assembled and analyzed the high-quality genomes of two wild tobacco species, *Nicotiana attenuata* and *Nicotiana obtusifolia*, to investigate the evolution of their adaptive traits and showed that an interplay of transposable element (TE) insertions and gene duplications contribute to this evolution. Using phylogenomic analyses, a deep analysis of nicotine, an important ecologically secondary metabolite, revealed that the involved pathway gradually evolved from two primary metabolic pathways followed by rapid acquisition of a root specific gene expression. Additionally, we showed that rapidly expended repetitive elements that were introduced after the Solanaceae whole genome duplication contributed to the high variability of genome sizes in *Nicotiana* and to the evolution of herbivory-induced signaling and defenses. Genome-wide analyses revealed that TE insertions promote expression divergence of duplicated genes which can lead to novel functions. Specially, transcription factor binding sites introduced by TE insertions have a high potential to coordinate the regulation of involved biosynthetic pathways and novel traits.

Author’s contributions:

S. Xu and I.T. Baldwin conceived and coordinated the project. K. Gase coordinated the sample collections for DNA and RNA sequencing and the submission of the genome to NCBI. K. Gase, H. Kuhl, and B. Timmermann coordinated the sequencing of the two genomes. H. Klaus and S. Xu assembled the genomes. S. Xu, T. Brockmöller, H. Tang, M. Stanke and E. Lyons annotated protein coding genes in the
genomes. T. Brockmöller and S. Xu performed functional annotation of the genes. P. Pandey and S.P. Pandey annotated smRNAs in *N. attenuata*. T. Brockmöller and S. Xu performed evolutionary genomics analyses to estimate species divergence times, to detect homologous groups, to identify and characterize gene duplication events and to estimate the age of the genes. T. Brockmöller performed gene-family expansion analysis. T. Brockmöller, Z. Ling and S. Xu analyzed RNA-seq and microarray data. S. Xu, T. Brockmöller, E. Gaquerel and A. Navarro-Quezada initiated and analyzed the evolution of nicotine biosynthesis, and analyzed the role of gene duplications and transposable elements in this process. C. Kreitzer, W. Zhou and K. Gase validated promoter region of nicotine biosynthesis genes using Sanger sequencing, S. Xu, E. Gaquerel and I.T. Baldwin wrote the manuscript.

In total, T. Brockmöller contributed 35% to the manuscript and the involved analyses.
3.2 Manuscript II

“Evolution of herbivore-induced early defense signaling was shaped by genome-wide duplications in Nicotiana”

(Zhou, W., Brockmöller, T., Ling, Z., Omdahl, A., Baldwin, IT., Xu, S.)

Published in eLife: November 4, 2016; DOI: 10.7554/eLife.19531

In manuscript II, we used comparative transcriptomics and network analyses to understand the recognition of herbivore attack and defense regulation in plants. We examined herbivore associated elicitors (HAE) that activate the early defense signaling in plants among closely related Nicotiana species and revealed a key gene co-expression network that is co-activated with HAE-induced jasmonic acid (JA) accumulations but independent from JA. Phylogenomic analysis showed a large gene retention rate within this key network after the whole-genome triplication event (WGT) shared among Solanaceae species. This might provide increased network complexity and robustness and can therefore have a positive effect on plant fitness.

Author’s contributions:

W. Zhou carried out the sampling, RNA preparation, performed qPCRs, cellular localization and VIGS experiments and analyzed the data. T. Brockmöller constructed phylogenetic trees and identified and characterized all gene duplications events for each gene, and analyzed and interpreted the data. Z. Ling performed RNA-seq mapping and quantification and assisted the RNA-seq data analysis. A. Omdahl performed motif analysis, and analyzed and interpreted the data. I. T. Baldwin drafted and revised the manuscript, contributed unpublished essential data and reagents. S. Xu conceived and coordinated the study, performed experiments, analyzed RNA-seq and microarray data and drafted and revised the manuscript.

In total, T. Brockmöller contributed 15% to this manuscript and the involved analyses.
3.3 Manuscript III

“Nicotiana attenuata Data Hub (NaDH): an integrative platform for exploring genomic, transcriptomic and metabolomic data in wild tobacco”

(Brockmöller, T., Ling, Z., Li, D., Gaquerel, E., Baldwin, I.T., Xu, S.)

Accepted in BMC Genomics: January 2017; DOI: 10.1186/s12864-016-3465-9

In manuscript III, we presented the interactive platform: Nicotiana attenuata Data Hub (NaDH). It combines genomic, phylogenetic and transcriptomic data (collected and generated in Manuscript I) including additional metabolomic datasets from mass spectrometry analysis to create a comprehensive toolbox for phylogenomic research and to understand the regulation and evolution of genes and linked traits. We introduced the different visualization methods and implemented tools of this web-based platform to examine gene to gene, gene to metabolite and metabolite to metabolite co-expression. With one case study, we introduced how researchers can use NaDH to reveal functions of genes and metabolites and their involved pathways.

Author’s contributions:

T. Brockmöller and S. Xu lead the project and designed the database and included tools. T. Brockmöller developed the homepage and database structure. T. Brockmöller and S. Xu generated the gene annotations and T. Brockmöller performed the phylogenetic analysis. Z. Ling provided the RNA-seq data. E. Gaquerel and D. Li performed the mass spectrometry analysis, annotated metabolites and inferred tissue-preferentiality of expression for metabolites. S. Xu, T. Brockmöller and I.T. Baldwin drafted and revised the manuscript. I.T. Baldwin provided the –omic data that is organized and used by NaDH.

In total, T. Brockmöller contributed 70% to the manuscript, the database development and the involved analyses.
4 Manuscript I
“Wild tobacco genomes reveal the evolution of prolific nicotine production”

Biological Sciences | Plant biology

Wild tobacco genomes reveal the evolution of nicotine biosynthesis

Shuqing Xu1,*, Thomas Brockmöller1,8, Aura Navarro-Quezada2, Heiner Kuhl3, Klaus Gase1, Zhihao Ling1, Wenwu Zhou1, Christoph Kreitzer1,4, Mario Stanke5, Haibao Tang6, Eric Lyons7, Priyanka Pandey8, Shree P. Pandey9, Bernd Timmermann3, Emmanuel Gaquerel1,*, and Ian T. Baldwin1*

Affiliations:
1Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Jena, Germany.
2Centre for Organismal Studies, University of Heidelberg, Heidelberg, Germany.
3Max Planck Institute for Molecular Genetics, Berlin, Germany.
4Veterinary University of Vienna, Vienna, Austria.
5Institute for Mathematics and Computer Science, Universität Greifswald, Greifswald, Germany.
6Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China.
7School of Plant Sciences, BIO5 Institute, CyVerse, University of Arizona, Tucson, USA.
8National Institute of Biomedical Genomics, Kalyani, West Bengal, India.
9Department of Biological Sciences, Indian Institute of Science Education and Research - Kolkat,
Mohanpur, West Bengal, India.
*Correspondence to Dr. Shuqing Xu (sxu@ice.mpg.de) or Dr. Emmanuel Gaquerel (emmanuel.gaquerel@cos.uni-heidelberg.de) or Prof. Ian T. Baldwin (baldwin@ice.mpg.de).
# these authors contributed equally.
Abstract

Nicotine, the signature alkaloid of *Nicotiana* species responsible for the addictive properties of human tobacco smoking, functions as a defensive neurotoxin against attacking herbivores. However, the evolution of the genetic features that contributed to the assembly of the nicotine biosynthetic pathway remains unknown. We sequenced and assembled genomes of two wild tobaccos, *Nicotiana attenuata* (2.5 Gb) and *N. obtusifolia* (1.5 Gb), two ecological models for investigating adaptive traits in nature. We show that after the Solanaceae whole genome triplication event, a repertoire of rapidly expanding transposable elements (TEs) bloated these *Nicotiana* genomes, promoted expression divergences among duplicated genes and contributed to the evolution of herbivory-induced signaling and defenses, including nicotine biosynthesis. The biosynthetic machinery that allows for nicotine synthesis in the roots evolved from the stepwise duplications of two ancient primary metabolic pathways: the polyamine and nicotinic acid dinucleotide (NAD) pathways. While the duplication of the former is shared among several Solanaceous genera which produce polyamine-derived tropane alkaloids, the innovation and efficient production of nicotine in the genus *Nicotiana* required lineage-specific duplications within the NAD pathway and the evolution of root-specific expression of the duplicated Solanaceae-specific ethylene response factor (ERF) that activates the expression of all nicotine biosynthetic genes. Furthermore, TE insertions that incorporated transcription factor binding motifs also likely contributed to the coordinated metabolic flux of the nicotine biosynthetic pathway. Together, these results provide evidence that TEs and gene duplications facilitated the emergence of a key metabolic innovation relevant to plant fitness.

Key words: *Nicotiana* genomes, genome-wide duplications, lineage-specific duplications, transposable elements, nicotine biosynthesis, ethylene response factor, expression divergence.
Significance Statement

Plants produce structurally diverse specialized metabolites, many of which have been exploited in medicine or as pest control agents, while some have been incorporated in our daily lives, such as nicotine. In nature, these metabolites serve complex functions for plants’ ecological adaptation to biotic and abiotic stresses. By analyzing two high-quality wild tobacco genomes, we provide an in-depth genomic study that directly associates genome evolution with the assembly and evolution of the nicotine biosynthetic machinery. These results demonstrate the importance of the interplay of gene duplications and transposable element insertions in the evolution of the multigenic biosynthetic pathways required of specialized metabolism and illuminates on how complex adaptive traits could evolve.
Introduction

The pyridine alkaloid nicotine, whose addictive properties are well-known to humans, is the signature compound of the genus *Nicotiana* (Solanaceae). In nature, nicotine is arguably one of the most broadly effective plant defense metabolites, in that it poisons acetylcholine receptors and is thereby toxic to all heterotrophs with neuromuscular junctions. Field studies using genetically-modified *N. attenuata* (coyote tobacco) plants, an annual wild diploid native to Western North America, have revealed that this toxin fulfills multifaceted ecological functions that contribute to plant fitness (1, 2). The strong transcriptional up-regulation of the nicotine biosynthetic machinery in roots in response to herbivore attack of the shoot combined with the active translocation and storage of this toxin provides *N. attenuata* plants with an inducible protection mechanism against a broad spectrum of herbivores (2). In addition, the transport and non-homogenous distribution of nicotine in the nectar of flowers within an inflorescence modifies the trap-lining behavior of humming bird pollinators to maximize outcrossing rates (3). These two facets of the ecological utility of nicotine result from the prolific production of this toxin which can accumulate up to 1% of the leaf dry mass in wild tobacco species (4). This prodigious biosynthetic ability is based on an efficient biosynthetic machinery composed of multiple genes co-expressed in roots (5, 6). In contrast to the deep knowledge on nicotine’s biosynthesis and ecological functions, the evolution of genomic features that facilitated the assembly of a pathway so critical for the survival of *Nicotiana* species has remained largely unknown.

Gene duplication and TE insertions continuously shape the evolutionary landscape of genomes and can affect the function of genes with adaptive consequences (7, 8). While whole-genome and local gene duplications provide the raw material for the evolution of novel traits, TE mobility can broadly remodel gene expression by redistributing transcription factor binding sites, shaping epigenetic marks and/or providing target sequences for small regulatory RNAs (7-11). Hence, the combination of gene duplications and TE activity is thought to facilitate the evolution of novel adaptive traits (8). However, the
details of this process, in particular its role in the evolution of metabolic complexity through the assembly
of novel multi-gene pathways, remains unclear.

Results and Discussion

Genome sequencing, assembly and annotation.

We sequenced and assembled the genome of *N. attenuata*, using 30x Illumina short reads, 4.5x
454 reads, and 10x PacBio single-molecule long reads. We assembled 2.37 Gb of sequences representing
92% of the expected genome size. We further generated a 50x optical map and a high-density linkage
map for super-scaffolding (Fig. S1 and S2), which anchored 825.8 Mb to 12 linkage groups and resulted
in a final assembly with a N50 contig equal to 90.4 kb and a scaffold size of 524.5 kb (Fig. S3). Likewise,
using ~50x Illumina short reads, we assembled the *N. obtusifolia* genome with a 59.5 kb and 134.1 kb
N50 contig and scaffold N50 size, respectively. The combined annotation pipeline integrating both hint-
guided AUGUSTUS and MAKER2 gene prediction pipeline predicted 33,449 gene models in the *N.
attenuata* genome. More than 71% of these genes models are fully supported by RNA-seq reads and
12,617 and 18,176 of these genes are orthologous to *Arabidopsis* and tomato genes, respectively.

To investigate the evolutionary history of the different *Nicotiana* genomes, we inferred 23,340
homologous groups using protein sequences from 11 published genomes (Table S1). A phylogenomic
analysis of the identified homologous groups demonstrated that *Nicotiana* species share a whole-genome
triplication (WGT) event with *Solanum* species, such as tomato, potato and *Petunia* (12), but not with
*Mimulus* (Figure 1 and Fig. S4-S7). At least 3,499 duplicated gene pairs originating from this WGT were
retained in both *Nicotiana* and *Solanum*. Among all retained duplicated gene pairs detected in *N.
attenuata* that did not further duplicate in this species, more than 53.7% showed expression divergence
(fold change greater than 2) in at least one tissue, indicating that these WGT-derived duplicated genes
may have evolved divergent functions through neofunctionalization or sub-functionalization.

Expansion of transposable elements in *Nicotiana*. 
Polyploidization is often associated with a burst of TE activity as a hypothesized consequence of “genomic shock” (13, 14). TEs, especially long terminal repeats (LTRs) are highly abundant in *Nicotiana* and account for 81.0% and 64.8% of the *N. attenuata* and *N. obtusifolia* genomes, representing significantly higher proportions than other sequenced Solanaceae genomes, such as tomato and potato (Figure 1). An analysis of the history of TE insertions revealed that all *Nicotiana* species experienced a recent wave of Gypsy retrotransposon expansion. However this expansion of Gypsy copies was less pronounced in *N. obtusifolia* compared to other *Nicotiana* species analyzed, which accounts for the smaller genome size of *N. obtusifolia*. A recent study showed that *Capsicum* species also experienced a large expansion of their Gypsy repertoire (15), albeit earlier than in *Nicotiana*, indicating that after WGT, the different Solanaceae lineages independently experienced the processes of Gypsy proliferation.

In addition to LTRs, miniature inverted–repeat transposable elements (MITEs), which are derived from truncated autonomous DNA transposons, may also play evolutionary roles. Although the size of MITEs is generally small, typically less than 600 bp, MITEs are often located adjacent to genes and are often transcriptionally active. As such, they have been hypothesized to contribute to the evolution of gene regulation (16, 17). In total, we identified 13 MITE families in the genome of *N. attenuata*, several of them having rapidly and specifically expanded in *Nicotiana* species (Figure 2a and b). Among these expanded MITE families, a Solanaceae-specific subgroup of the Tc1/Mariner defined by DTT-NIC1 is the most abundant. By analyzing insertion positions of this subgroup, we found that DTT-NIC1 copies, similar to other DNA transposons, are significantly enriched within a 1 kb region upstream of the genes in *N. attenuata* (Figure 2c). Analyses on the herbivory-induced conserved transcriptomic responses in *Nicotiana* further showed that insertions of DTT-NIC1 are significantly enriched within the 1 kb upstream region of herbivory-induced early defense signaling genes in *N. attenuata*, and may have contributed to the recruitment of genes into the induced defense signaling network by introducing WRKY transcription factor binding sites (18).
Innovations in metabolic and signaling network architecture are thought to result from the rapid rewiring of tissue-level gene expression patterns following duplication events (19, 20). To examine this inference, we compared the genome-wide expression divergence between duplicated gene pairs and analyzed the effects of DTT-NIC1 insertions into 1 kb upstream regions of each member of the gene pairs. Insertions of the DTT-NIC1 family were associated with significant divergences in expression and tissue specificity between duplicated genes (Figure 2d), consistent with the hypothesis that the expansion of this TE family was a critical determinant of genome-wide re-wirings of gene regulation occurring post-duplication in these *Nicotiana* species.

**Evolution of nicotine biosynthesis.**

To further understand the role of gene duplication and TE insertions on the evolution of *Nicotiana* adaptive traits, we reconstructed the evolutionary history of the nicotine biosynthesis pathway, a key defensive innovation of the *Nicotiana* genus. Nicotine biosynthesis is restricted to the roots and involves the synthesis of a pyridine ring and a pyrrolidine ring which are coupled most likely via the action of genes coding for an isoflavone reductase-like protein, called A622, and the berberine bridge enzyme-like (BBL) enzymes (21, 22) (Figure 3A). Phylogenomic analyses revealed that genes involved in the biosynthesis of the pyridine and pyrrolidine rings evolved from the duplication of two primary metabolic pathways that are ancient across all plant lineages: the nicotinamide adenine dinucleotide (NAD) cofactor and polyamine metabolism pathways, respectively (Figure 3a).

However, the timing and mode of duplications of these two pathways differ and reflect the expansion and recruitment of gene sets required for the diversification of alkaloid metabolism in the Solanaceae. Duplications that gave rise to the branch extension of the polyamine pathway required for the biosynthesis of the signature alkaloids of Solanaceae and Convolvulaceae (e.g., tropane in many genera and nicotine in *Nicotiana*) are shared among *Nicotiana*, *Solanum*, and *Petunia* with individual gene members recruited from the Solanaceae WGT or earlier duplication events. Genes encoding ornithine
decarboxylase (ODC2) and N-putrescine methyltransferase (PMT) duplicated prior to the shared
Solanaceae WGT from their ancestral copies in polyamine metabolism, ODC1 and spermidine synthase
(SPDS), respectively (Fig. S8 and S9). While ODC2 likely retained its ancestral enzymatic function, PMT
(derived from SPDS) acquired the capacity to methylate putrescine to form N-methyl-putrescine through
neofunctionalization (23). The N-methylputrescine oxidase (MPO) from the polyamine metabolism
pathway evolved from diamine oxidase (DAO)(24) through whole genome multiplication. Both copies
were retained in Nicotiana, Solanum and Petunia (Fig. S10), presumably to sustain the flux of N-methyl-
Δ1-pyrroline required for alkaloid biosynthesis. Duplication patterns of ODC, PMT and MPO therefore
support the ancient origin of the ornithine-derived N-methyl-Δ1-pyrroline, which is utilized as a
common building block for the biosynthesis of most alkaloid groups in the Solanaceae and
Convulvulaceae.

In contrast to the relatively ancient origin of pyrroline ring biosynthesis, duplications of the
NAD pathway genes, encoding aspartate oxidase (AO) and quinolinic acid phosphoribosyl transferase
(QPT), responsible for pyridine ring biosynthesis are Nicotiana-specific and likely occurred through local
duplication events (Fig. S11 and S12). BBLs are thought to be involved in the late oxidation step in
nicotine biosynthesis that couples the pyridine and pyrroline rings, and therefore constitute a key
innovation in the Nicotiana-specific synthesis of pyridine alkaloids. BBLs exhibiting clear root expression
specificity and likely evolved through neofunctionalization after gene duplications (Fig. S13).

Tissue-level RNA-seq transcriptome analyses in N. attenuata confirmed that while ancestral
copies exhibit diverse expression patterns among different tissues, all of the duplicated gene copies
recruited for nicotine biosynthesis are specifically expressed in roots (Figure 3b) and also specifically
transcriptionally up-regulated in response to herbivory via the jasmonate signaling pathway (25).
Experimental work has shown that the transcription factors of the ethylene response factor (ERF189)
subfamily IX and MYC2, play central roles in the up-regulation of nicotine genes(5). Analyzing the
evolutionary history of MYC2 revealed that this gene duplicated at the base of the Solanaceae via
3+
1
8;

tandemly duplicated independently in each lineage, but only evolved root-specific expressions in the
diploid species of *Nicotiana* (Figure 4). Because of its essential role in regulating the expression of all
nicotine genes (6), the acquisition of a root-specific expression by *ERF189* in the ancestor of *Nicotiana*
species likely played a critical role for the coordinated root expression of nicotine biosynthesis genes.

Regulation of nicotine genes by *MYC2* and *ERF189* relies in part on the presence of two
transcription factor binding sites, the GCC and G-box elements in their promoter regions (5, 6, 27).
Nicotine biosynthetic genes harbor more than twice the frequency of GCC and G-box elements in their 2
kb upstream region than do their ancestral copies (Figure 3b), consistent with the hypothesis that the
accumulation of GCC and G-box elements in promoter regions contributed to the evolution of the
coordinated transcriptional regulation required for high-flux nicotine biosynthesis. Investigating the origin
of GCC and G-box motifs in upstream regions of 10 nicotine biosynthesis genes showed that at least 34.8%
and 29.0% of GCC and G-box motifs, respectively (Figure 3c), are likely derived from TE insertions.
While it is unclear whether all of these TE-derived GCC and G-box motifs are involved in regulating the
expression of nicotine genes, some likely are. For example, in the case of *PMT1*, a previous experimental
study revealed that the 650 bp upstream region, which specifically contained additional TE-derived GCC
motifs, had a much larger capacity to drive the expression of a reporter gene in *Nicotiana* roots than did
the 111 bp upstream region that lacked these motifs (28). Furthermore, all GCC and G-box motifs within
2kb 5' region of the MPO1 that is likely under control of *ERF189* (24) are derived from TEs (Figure 3c).

The mechanisms of genome organizational evolution, such as genome-wide duplications and TE
expansions, facilitated the evolution of several aspects of the anti-herbivore defense arsenal including a
key metabolic innovation in *Nicotiana* species. These results are consistent with the hypothesis that TEs,
which have often been considered as ‘junk’ DNA, can be important orchestrators of the gene expression
remodeling that is required for the evolution of adaptive traits. Since native Nicotiana species do not
survive in nature without the ability to produce large quantities of nicotine to ward off attackers, it is
likely that this ‘junk’ has inspired innovation essential for their survival (29).

Conclusion

We sequenced, de novo assembled, and annotated the two genomes of two species of Nicotiana, the genus of which is scientifically and economically important. The fully annotated gene models, transposable elements, smRNAs and transcriptomic atlas (SI Appendix Table S1-S5, Dataset S1-S9) of N. attenuata enable comparative analysis to illuminate the evolution of specialized metabolites and novel adaptive traits in Solanaceae as demonstrated by our in-depth genomic analysis on the evolution of nicotine biosynthesis.

Materials and Methods

Plant material and DNA preparation

Plants were grown as previously described (30). The genomic DNA sequenced by 454 and Illumina HiSeq2000 technologies was isolated from late rosette-stage plants using the CTAB-method(31).
The two N. attenuata DNA plants used for this extraction were from a 30th generation inbred line, referred to as the UT accession, which originated from a 1996 collection from a native population in Washington County, Utah, USA (30). N. obliquifolia DNA was obtained from a single plant of the first inbred generation derived from seeds collected from a native population in 2004 at the Lytle Ranch Preserve, Saint George, Utah, USA. High molecular weight genomic DNA used to generate the optical map of N. attenuata was isolated using a nuclei based protocol (32) from approximately one hundred freshly harvested N. attenuata plants (harvested 29 days post germination) of the same inbred generation and origin as used for the short-read sequencing.
To reduce the potential effects of secondary metabolites on single molecular sequencing, the plant material used for PacBio sequencing was from a cross of two isogenic *N. attenuata* transgenic lines (mother: *ir-aoc*, line A-07-457-1, which was transformed with pRES5AOC [GenBank KX011463] and is impaired in JA biosynthesis; father: *rgPPS*, line A-07-230-5, which was transformed with pRES5GGPPS [GenBank KX011462] and is impaired in the synthesis of the abundant 17-
hydroxygeranyllinalool diterpene glycosides), both generated from the 22nd inbred generation of the same origin as the *N. attenuata* plants described above (30). Genomic DNA was isolated from approximately one hundred young plants (harvested 29 days post germination) by Amplicon Express (http://ampliconexpress.com) according to a proprietary protocol.

Genome sequencing and assembly

For *N. attenuata*, the Illumina HiSeq2000 system was used to generate a high coverage whole genome shotgun sequencing (WGS) of the genome based on short reads (2 × 100 bp or 2 × 120 bp). Different paired-end libraries were constructed using the Illumina TruSeq DNA sample preparation kit v2. The fragment size distribution maxima were observed at 180, 250, 600 and 950 bp. Additionally, two mate-pair libraries were constructed using Illumina mate-pair library preparation kit v2, which had their maxima at 5,500 and 20,000 bp of the fragment size frequency distribution, respectively. A lower genome-wide coverage of long reads (median read length 780 bp) was generated by the Roche/454 GS FLX (+) pyro-sequencing technology using Roche rapid library prep kit v2. For *N. obtusifolia*, two paired-end libraries and a single mate-pair library were constructed using the same material. The two paired-end libraries had fragments size distribution maxima at 480 and 1050 bp, respectively. The mate-pair library had a maximum at 3500 bp. The 20 kb mate-pair libraries were constructed at Eurofins/MWG using the Cre-recombinase circularization approach from Roche (Roche Diagnostics GmbH, Mannheim, Germany). These were both sequenced with 454 technology and Illumina HiSeq2000 (by removing Roche adaptor sequences and replacing them by Illumina adaptor sequences). The PacBio reads were sequenced at the Cold Spring Harbor Laboratory.
The overall assembly workflow for *N. attenuata* is shown as Fig. S3. All paired-end reads from the sequenced libraries were assembled using the Celera Assembler (CA7) with a minimum read length cut-off at 64bp. Preliminary tests showed that single end or short reads did not improve the assemblies, but increased calculation time significantly. In total, 86.4 Gb short read data were assembled. The expected genome size of *N. attenuata* is of 2.54 Gb based on coverage of the “larger than N50 length unitigs”, similar to the estimation (1C=2.5pg) from the flow cytometry analysis. We used the SSPACE v2 scaffold to further improve scaffolding using the mate-pair data and filled gaps using GapCloser v1.12 (33). After manual inspections, we found that certain neighboring contigs in the scaffolds still contained overlaps, which might be due to the assembly process from CA7 that places copies of repeat sequence at the end of contigs or due to issues in Gapcloser v1.12 that leave open some closable gaps. To close these gaps between overlapped contigs, we compared neighboring contigs using BLASTN (min. identity 95%/min. length 43) and then joined overlapping contigs with custom scripts.

The assembly scaffolds from short reads were used for gap filling and further scaffolding using PBJelly (v15.8.24)(34) with ~10x PacBio reads (N50=14.9 kb, max read length =48.9 kb), which resulted in a 2.17 Gb genome assembly. While PBjelly only increased the N50 scaffold size from 176 kb to 202 kb, it significantly increased the N50 contig size from 67 kb to 90 kb. Because PacBio reads contain about 12-15% errors, we performed an additional correction step using short reads with PILON(35). In total, 98.2% of the draft assembly was confirmed by short reads and ~1.3 Mb sequences were corrected by PILON. The PILON corrected assembly was further mapped to the 10x PacBio reads using BLASR and used SSPACE-longreads for the second round scaffolding, which increased the N50 scaffold size to 292 kb.

To assemble the *N. obtusifolia* genome, we employed a hybrid strategy in which we first assembled all short reads by a ‘de Bruijn graph’ assembler using idba-ud v1.1.1 (36), and then assembled the locally re-assembled contigs and a subset of the short read data by an ‘OLC’ long read assembler.
using CA7. Scaffolding and gap filling were carried out using SSPACE v2 using mate-pair data in a
similar manner to the \textit{N. attenuata} assembly.

Analyses of 248 conserved core eukaryotic genes using the CEGMA(37) pipeline indicated that
both \textit{N. attenuata} and \textit{N. obtusifolia} assemblies were only slightly less complete in full-length gene
contents than that of the tomato genome (86.7\%) and similar to the assembly of the potato genome
(83.9\%) (Table S2).

\textbf{Annotation of transposable elements}

\textit{De novo} annotation of repeated elements was performed with RepeatModeler version open-4-0-5
with the parameters \texttt{-engine ncbi}. We identified 667 consensus repeat sequences (1.3 Mb total size) in
the \textit{N. attenuata} genome. To classify these consensus repeat sequences, additional annotation using
TEclass (38) was applied for repeats that were not classified by RepeatModeler. Among all identified
repeats, LTRs, DNA transposons and LINEs contributed most, representing 47.5\%, 28.3\% and 9.3\%,
respectively. The annotated repeats were used for masking repeat sequences using RepeatMasker (open-
4.0.5) using parameter \texttt{"-e nebi -norna"}. We further re-annotated transposable elements using the \textit{N.}
\textit{attenuata} repeat library for two \textit{Nicotiana} additional genomes: \textit{N. sylvestris} and \textit{N. tomentosiformis} (39).
To make the results comparable, we used the same approach to \textit{de novo} identify the TE library of \textit{S.}
\textit{lycopersicum} (26) and \textit{S. tuberosum} (40) genomes.

MITEs in \textit{Nicotiana} were annotated in two steps. First, MITE-Hunter (41) was used to find MITE
families in the \textit{N. attenuata} genome using default parameters, except \texttt{"-P 0.2"}. Following the manual of
MITE-Hunter, the identified MITE candidate families were first subjected coverage evaluation using
TARGeT. The output results were manually inspected and only the MITE families that showed even
distribution of coverage were selected. Then these selected candidate MITE families were manually
checked for their terminal inverted repeats (TIR) and target site duplications (TSD). In total, 15 MITE
families were initially identified. We then assigned these 15 MITE families to different super-families and
classes based on sequence homology to a P-MITE database. Two MITE families that showed no
homology to any known MITE sequences were excluded from the downstream analysis. Second, using
these 13 MITE consensus sequences as a library, we identified the copy number of each MITE family
using RepeatMasker with parameters “-nolow -no_is -s -cutoff 250”. A complete MITE sequence was
defined as being no more than 3 bp shorter than the representative sequence. The multiple sequence
alignment and neighbor joining tree construction of the DTT-NIC1 family were performed using
ClustalW.

**Annotation of protein-coding genes**

The *N. attenuata* genome was annotated using the *Nicotiana* Genome Annotation (NGA) pipeline,
which employs both hint-guided (hg) Augustus (v. 2.7) (42) (hg-Augustus) and MAKER2 (v.2.28)(43),
gene prediction pipelines based on genome release v1.0. For hg-Augustus gene annotation, the HMM
gene model was specifically trained for *N. attenuata* using RNA-seq data from major plant tissues, and
gene models were predicted using unmasked genome sequences. The repeat regions were here given less
probability to be predicted as a gene, in particular when RNA-seq evidence was missing. For MAKER2
annotation pipeline, we integrated evidence of multiple protein codings from three sources: *ab initio* gene
predictions, transcript evidence and protein homolog evidence. The input evidences for MAKER2 were: 1)
*ab initio* gene predictors (GeneMark and Augustus) that were each trained with full-length transcripts; 2)
Trinity (v. r20131110) assembled transcripts using RNA-seq data from major tissues; 3) UNIREF90 plant
proteins (mapped using genewise, v. wise2-4-1); 4) six high quality plant proteomes (tomato, potato,
grape, *Arabidopsis*, *Populus* and rice). The MAKER2 annotation pipeline was run on the repeat masked *N.
*attenuata* genome.

The predicted gene models from hg-Augustus were filtered based on their repeats contents. All
genes with repeats occupying more than 50% of the gene length were removed, and genes were retained if
less than 10% of their sequence matched to repeats. For genes that contained repeats occupying 10-50%
of their entire gene length, we performed an additional search of the plant Refseq database using
BLASTX. If the gene matched with a non-repeats homolog (e-value greater than 1e-5 and bit score greater than 200) from the plant refseq database, the predicted gene models were retained for downstream analysis. In total, 35,737 gene models from hg-Augustus predictions were retained. For MAKER2 predicted gene models, in addition to the filtering based on repeat content as described above, the gene models that had low evidence support were removed from downstream analysis (eAED <= 0.45, this cutoff was set after manual inspections on the predicted gene models). In total 33,274 gene models from MAKER2 prediction were retained. The predicted gene models from hg-Augustus and MAKER2 pipelines were then combined and overlapping gene structures were removed. After manual inspections, we found that hg-Augustus outperformed MAKER prediction when the pipelines predicted different gene structures for the same gene. Therefore, when the two pipelines predicted different gene structures for the same genome region, we retained only the hg-Augustus predicted gene models. After merging the predicted models from both hg-Augustus and MAKER2 predictions, genes with pre-mature stop codons were considered as pseudogenes and were removed from the downstream analysis. The predicted gene models were then transferred to *N. attenuata* genome release v2.0 using a custom script which first identified homologous regions using BLAST and then predicted gene structure using GeneWise. This finally resulted in 33,449 high quality gene models in the final *N. attenuata* genome, of which 74.9% were supported by at least 50 RNA-seq reads. Because gene models within a genus are usually conserved, we annotated *N. obtusifolia* and two other publicly available diploid *Nicotiana* genomes, *N. sylvestris* and *N. tomentosiformis* (39), using a homology-based annotation pipeline based on all predicted *N. attenuata* protein coding gene models. Using all predicted gene models from *N. attenuata*, we predicted protein coding gene models in *N. obtusifolia*, *N. tomentosiformis* and *N. sylvestris* using a homolog-based approach.

**RNA sequencing and data analysis**

RNA was isolated from plants of the same origin and inbred generation as described above for *N. attenuata* DNA isolation. RNA was isolated using TRIZOL® (Thermo Fisher Scientific) according to the
instructions of the manufacturer. DNA was removed from all RNA preparations using TURBOD<sup>TM</sup> DNase (Thermo Fisher Scientific) according to the manufacturer’s protocol. In total, twenty one RNA-seq libraries from different plant tissues and the same tissues under different biotic and abiotic stress treatments were first enriched for RNAs with poly-A tails and then used for RNA-seq library construction with Illumina's TruSeq RNA sample preparation kits. The insertion sizes of the libraries are approximately 200 bp. All RNA-seq libraries were sequenced using Illumina 2000 HiSeq platform with read lengths of 50 or 101 bp and resulted in 793,785,373 paired-end Illumina reads.

The raw sequence reads were trimmed using AdapterRemoval (v1.1) with parameters “--collapse --trimns --trimqualities 2 --minlength 36”. The trimmed reads were then aligned to the <i>N. attenuata</i> genome assembly (v 2.0) using TopHat2 (v2.1.0) (44), with maximum and minimum intron size set to 50,000 and 41 base pairs (bp), respectively, estimated from the <i>N. attenuata</i> genome annotation.

To estimate the expression level of assembled genes and transcripts, all trimmed RNA-seq reads were mapped to the assembled transcripts using RSEM (v1.2.20) (45). Transcripts per million (TPM) was calculated for each transcript and gene. To exclude low-expressed genes and transcripts, only genes with TPM greater than five in at least one sample were considered as expressed. Similarly, only the transcripts with TPM greater than one in at least one sample were considered as expressed.

**Comparative genomic analysis**

We assigned genes to homologous groups (HGs) using a similarity-based method. For this, we used all genes that were predicted from the 11 genomes, listed in Table 1. All-vs-all BLAST analysis was used to compare the sequence similarity of all protein coding genes, and the results were filtered based on the following criteria: e-value less than 1e-20; match length greater than 60 amino acids; sequence coverage greater than 60% and identity greater than 50%. All remaining blast results were then clustered into HGs using a Markov clustering algorithm (MLC) (46).

We constructed a phylogenetic tree for all identified HGs using an in-house developed pipeline (SI appendix).
Evolution of nicotine biosynthesis and GCC and G-box transcription factor binding sites

Genes previously characterized as being involved in nicotine biosynthesis were retrieved from the literature and sequences were downloaded from NCBI. Phylogenetic trees for each nicotine biosynthesis gene were constructed as described above. Sequence alignment and tree structures were then manually inspected. Duplication events of nicotine biosynthetic genes were inferred from the phylogenetic tree structures as well as, when possible, from manually checking syntenic information from the tomato and potato genomes.

We extracted the GCC and G-box motif matrix from the literature (5, 27), and used FIMO (47) to detect the occurrence of these two motifs within the 2kb upstream regions of both nicotine biosynthesis genes and of their ancestral/non-root specific copies. Only the motifs with e-values less than 1e-3 were considered. Manually inspecting the positions of the annotated motif regions revealed that several motifs overlapped with annotations of TEs, such as in the upstream regions of MPO and PMT, indicating that some of these motifs may be derived from TE sequences. To test this hypothesis, we first searched GCC and G-box motif sequences within the consensus TE sequences. Overall, GCC and G-box motifs could be found in more than 54% of these TE consensus sequences. The number of GCC box and G-box motifs per kilo-base sequences in the TE consensus sequences were as high as 19 and 28, respectively. Permutation tests by randomly shuffling the positions of GCC and G-boxes 1000 times in the N. attenuata genome and then comparing with actual TE locations further revealed that these two motifs were significantly enriched in TE regions ($p < 0.001$). These data reveal that many TEs contain the GCC and G-box motif sequences. Next, we performed additional analyses in order to calculate the number of the GCC and G-box motifs which were derived from TE insertions that were located within the upstream regions of nicotine biosynthesis genes. For this, we extracted 150 bp sequences that included left and right flanking sequences and the motif sequence in the middle and compared these with the RepeatMasker annotated TE sequences in the N. attenuata genome using YASS (48), a tool designed to search for diverged sequences. To reduce false positives, only the matches that contained the expected motif sequences and had an e-
value lower than 1e-5 were considered. Note that the number of GCC and G-box motifs in the nicotine biosynthesis genes that derive from TEs estimated by this approach is likely highly conservative, because this method fails to identify the corresponding homologous sequences in cases where the motif sequences and their flanking regions have diverged significantly from their ancestral TE sequences.
References:


Acknowledgements:

We acknowledge the following sources for funding: Swiss National Science Foundation (No. PEBZP3-142886 to SX), the Marie Curie Intra-European Fellowship (IEF) (No. 328935 to SX), European Research Council advanced grant ClockworkGreen (No. 293926 to ITB), DFG Exzellenzinitiative II to the University of Heidelberg (EG and ANQ) and the Max Planck Society, which provided all of the funds for the sequencing. The CoGe platform (www.genomevolution.org) is supported by NSF (IOS – 1339156 and IOS – 1444490). We thank members from the Department of Molecular Ecology for assistance with the manual curation of gene models and for scientific discussions, Dr. Sang-Gyu Kim and Dr. Matthias Erb for help with RNA sample collections, Dr. Alex Hastie from BioNano Genomics for assembling of the BioNano optical map. We thank Prof. Jonathan Gershenzon, Dr. Ewald Grosse-Wilde and Nicolas Arning for their constructive comments and suggestions on an earlier version of the manuscript.

Author contributions:

SX and ITB conceived and coordinated the project. KG coordinated sample collections for DNA and RNA sequencing and the submission of the genome to NCBI. KG, HK, and BT coordinated the sequencing of the two genomes. HK and SX assembled the genomes. SX, TB, HT, MS and EL annotated protein coding genes in the genomes. PP and SPP annotated smRNAs in *N. attenuata*. TB and SX performed comparative genomic analysis. TB, ZL and SX analyzed RNA-seq and microarray data. SX, TB, EG and ANQ initiated and analyzed the evolution of nicotine biosynthesis and transposable elements. CK, WZ and KG validated promoter region of nicotine biosynthesis genes using Sanger sequencing. SX, EG and ITB wrote the manuscript.

The authors declare no competing financial interest.
528 Figure legends:

![Figure 1](image)

**Figure 1.** Whole genome triplication (WGT) in *Nicotiana* genomes is shared with other Solanaceae species but the Gypsy retrotransposons expansions are *Nicotiana*-specific.

(a) *Nicotiana* genomes share the WGT with other Solanaceae species. Left panel depicts the shared WGT event between *Nicotiana* and *Solanum* as revealed by the structure of the phylogenetic tree of triplicated gene families in *Nicotiana* and *Solanum*. Red and yellow bars represent the percentage of triplicated and duplication events shared between *Nicotiana* and *Solanum*, respectively. Right panel shows the phylogenetic tree of 11 plant species and different colored stars indicate previously characterized whole genome multiplication events. (b) Expansion of Gypsy transposable elements contributes substantially to genome size evolution in *Nicotiana*. Left panel shows the genomic content (in Gb) of repetitive versus non-repetitive sequences in the 11 plant genomes. Black and grey bars indicate non-repetitive sequences, whereas other colors indicate repetitive sequences. The right insert visualizes the expansion history of LTR retrotransposons in four *Nicotiana* genomes in comparison to tomato. X-axis (number of substitutions per site) refers to the divergence of a LTR from its closest paralog in the genome, with smaller numbers indicating more recent duplication events.
Figure 2. Expansion of transposable elements of the family DTT-NIC1 increased genome-wide gene expression divergence among duplicated gene pairs in *Nicotiana*.

(a) Copy number of the six most abundant *Nicotiana* MITE families of transposable elements in *Nicotiana* and *Solanum*. Each bar depicts the total number of copies in each species for the six main MITE transposable elements (TEs). MITE families are visualized by different colours: light blue, DTA *(hAT)*; green, DTH *(PIF/ Harbinger)*; red, DTT *(Tc1/Mariner)*. DTT-NIC1 from the Tc1/Mariner family is the most abundant all MITE TEs. *Nicotiana* species: Na, *N. attenuata*; No, *N. obliquifolia*; Ns, *N. sylvestris*; Nt, *N. tomentosiformis*. *Solanum* species: Sl, *S. lycopersicum*; St, *S. tuberosum*. (b) Expansion of the DTT-NIC1 family in *Nicotiana* species. Neighbor joining (NJ) tree of the DTT-NIC1 family in *N. attenuata*, tomato and potato. The shaded clade highlights the pronounced expansion of DTT-NIC1 in *N. attenuata*. (c) DTT-NIC1 insertions are enriched in the upstream regions of coding sequences. The line indicates the percentage of genes, among all predicted protein coding genes, that contain DTT-NIC1 insertions within a given 500 bp sliding window. (d) Insertions of DTT-NIC1 within
the 1 kb upstream region of duplicated genes increased tissue-level gene expression divergence (Wilcoxon rank sum test). Left and right panels are violin plots of the divergences between duplicated gene pairs at expression and tissue specificity levels, respectively. Red bars indicate duplicated pairs, of which one copy has at least one DTT-NIC1 insertion and the other does not. Blue bars indicate duplicated pairs, both of which lack DTT-NIC1 insertions. The width of the probability density in the violin plots along the bars correspond to the number of duplicate gene pairs.
Figure 3. Prolific nicotine production evolved from the duplication of two primary metabolic pathways and its coordinated transcriptional regulation was likely facilitated by transposon-derived transcription factor binding site insertions.

(a) *Nicotiana*-specific nicotine biosynthesis originates from step-wise duplications of two primary metabolic pathways. The upper panel depicts the metabolic organization (brightly colored and dashed line outlined branches) and evolution of nicotine biosynthesis via pathway and single gene duplications in *Nicotiana*. Light green and light blue branches on the side indicate the two ancient gene modules with housekeeping functions in plants corresponding to the NAD cofactor and polyamine pathways. Different gene duplication types are indicated by arrows annotated as follows: NSD, *Nicotiana*-specific duplications; WGT, whole genome triplication in Solanaceae; Prior-WGT, gene duplication occurring
prior to WGT. *Nicotiana* QS did not duplicate but experienced an increase in root expression compared to
its tomato homolog. Lower panel: phylgenomics view of grape, tomato and *N. attenuata* gene sets
highlighting the gradual assembly of the nicotine biosynthetic pathway. (b) Acquisition of transcription
factor binding motifs and root-specific expression evolution of nicotine biosynthesis genes. Heatmaps
derpect the scaled expression of nicotine biosynthetic genes and their ancestral copies or closest paralogs
across six distinct tissues. Red and green signify high and low expression, respectively. A622 likely
neofunctionalized without being duplicated in *solanaceous* species. TPM: transcript per million. Nicotine
biosynthetic genes’ root specific expression and dramatic transcriptional up-regulation during insect
herbivory is coordinated by the action of MYC2 and ERF transcription factors which target G- and GCC-
type boxes in the promoters, respectively. Numbers of GCC and G-box motifs detected within 2 kb
upstream region of nicotine biosynthetic genes and their ancestral copies are represented using specific
color gradients. GCC motifs derived from TE insertion in the gene upstream region are shown as blue
lines. (c) Many GCC and G-box motifs from nicotine biosynthesis genes are likely derived from TE. Each
row depicts the motif and TE annotation of the 2 kb upstream region of an individual nicotine
biosynthesis gene. The predicted GCC and G-box motifs are shown in red and pink small boxes,
respectively. The regions that were annotated as TE from RepeatMasker are shown in rectangle with two
different colors. Light blue: LTR; green: non-LTR. The motifs sequences and their 150 bp flanking region
showed significant homology (E-value less than 1e-5) to annotated TE sequences in *N. attenuata* are
shown in dashed lines.
Figure 4. Evolution of ERF189 in Nicotiana. (a) Phylogenetic tree of ERF189 among different plants. The numbers on the branches denote approximate Bayes branch support values. Red arrows indicate the tandem duplications that were identified based on syntenic information. (b) Evolutionary model of the
ERF189 gene cluster based on both phylogenetic relationship among homolog genes and their syntenic information in *P. axillaris* (12), tomato (26) and *N. obtusifolia*. Lineage specific duplications of ERFs are shown in blue (*Petunia*), light blue (*Solanum*) and red (*Nicotiana*). The expression profile of ERF189 and its orthologous gene in tomato were retrieved from *N. attenuata* data hub (nadh.ice.mpg.de) and tomato eFP browser (http://bar.utoronto.ca/efp_tomato/cgi-bin/efpWeb.cgi), respectively. The asterisk indicates the gene NIOBTv3_g42088 is likely a pseudogene in *N. obtusifolia*. 
Supplementary Information Appendix

1. Sequencing, assembly and map integration of the Nicotiana attenuata and N. obtusifolia genomes .............33
   1.1 Raw data processing .........................................................................................................................33
   1.2 Generating the optical map ..............................................................................................................33
   1.3 Construction of a high-density linkage map ...................................................................................34
   1.4 Anchor scaffolds to pseudomolecules ...............................................................................................35
   1.5 Quality assessment and validation of assembly ...............................................................................35

2. Genome annotation ..................................................................................................................................36
   2.1 Annotation of repeats and LTR insertion time estimation in Nicotiana ........................................36
   2.2 Analyzing DTT-NIC1 insertions in Nicotiana genomes ................................................................37
   2.3 Annotation of miRNA, tRNA and rRNA ..........................................................................................37
   2.4 Gene and transcript expression analysis ..........................................................................................39
   2.5 Expression of transposable elements ..............................................................................................40
   2.6 Functional annotation of protein coding genes ...............................................................................41
   2.7 Functional annotation of protein domains ......................................................................................41
   2.8 Annotation of transcription factor and protein kinase genes ..........................................................42

3. Evolutionary genomics analyses ..............................................................................................................42
   3.1 Homolog group identification ..........................................................................................................42
   3.2 Detection of gene duplication events in HG ....................................................................................43
   3.3 Confirmation of the whole-genome triplication in Nicotiana ............................................................45
   3.4 Estimation of species divergence times ...........................................................................................46
   3.5 Identification of lineage-specific gene family expansion ...............................................................47

4. Evolution of nicotine biosynthesis .........................................................................................................49
   4.1 Identification and reconstruction of the evolution history of nicotine biosynthesis genes ............49
   4.2 Validation of the promoter sequences of nicotine biosynthesis genes and their ancestral copies ....49
   4.3 Prediction of TE-derived putative miRNA target sites into regulatory region of nicotine biosynthesis genes ..50
   4.4 Evolution of root-specific expression of nicotine biosynthesis genes ..........................................50
   4.5 Data access .......................................................................................................................................51

5. Supplemental Tables ...............................................................................................................................53
6. Supplemental Figures ..............................................................................................................................62
7. Captions for supplementary datasets S1 to S9 ....................................................................................96
8. References ...............................................................................................................................................98
1. Sequencing, assembly and map integration of the *Nicotiana attenuata* and *N. obtusifolia* genomes

1.1 Raw data processing

We filtered and trimmed the Illumina HiSeq2000 whole-genome shotgun raw sequences of *N. attenuata* and *N. obtusifolia* to obtain high-quality non-duplicated read pairs prior to genome assembly. This was achieved by a custom script that extracted only the largest reads, longer than 32 bp, and which contained no bases with a Phred quality score lower than 11. We compared the first 32 bp of each read in a pair to those of other read pairs and retained only one pair if the same sequence was found more than once (deduplication). Roche/454 long reads for *N. attenuata* were processed by the sffToCA script provided by the Celera Assembler v7 pipeline (CA7). Table S3 provides details of the filtered data used for the genome assemblies.

1.2 Generating the optical map

The optical maps of *N. attenuata* obtained from BioNano Genomics were used to improve the assembly quality. High molecular weight genomic DNA was isolated from fresh *N. attenuata* tissues using the protocol similar to that of VanBuren et al.(1). Briefly, around 5g leaves were collected from young plants and fixed with formaldehyde. After being homogenized in isolation buffer, filtrating washing treatment was performed. The nuclei were purified on percoll cushions and washed extensively and finally embedded in low melting agarose at different dilutions. The DNA plugs were treated with a lysis buffer containing detergent, proteinase K and β-mercaptoethanol (BME). In total, 140 Gb of data (>100 kb) were collected representing ~55× genome coverage with a molecule N50 length of 250 kb (Fig. S1). Molecules were *de novo* assembled as previously described (2). The optical map finally assembled consists of 2.2 Gb with an N50 size of 1.4 Mb.

The optical map final assembly was then used to anchor sequence scaffolds using the sewing machine pipeline (3). Overall 81.7% of sequence assembly can be mapped to the optical map and 84.7% of the optical map can be aligned to sequence assembly. The super scaffolding was performed using
parameter ‘--f_con 12 --f_algn 40 --s_con 10 --s_algn --T 1e-8’ after manual inspection of the quality of
the scaffolds. The super scaffolding using the optical map generated a genome assembly consisting of
39,115 scaffolds with N50 to 358.4 kb (total size 2.4 Gb).

1.3 Construction of a high-density linkage map

A high-density linkage map was constructed using the genotyping by sequencing (GBS) method
on 256 individuals from an advanced intercross recombinant inbred line population (AI-RIL). The
establishment of the AI-RIL and detailed procedures for DNA isolation and sequencing is described in
detail elsewhere (Zhou et al. submitted). In total, 1.2 billion paired-end clean reads were obtained from
256 samples (average = 4.7 million) after quality control and adapter trimming using AdapterRemoval (4)
with parameter “--minquality 30 --minlength 36”. All reads were then mapped to the draft genome of *N.
attenuata*(release v1.0) using Bowtie 2 (version 2.2.5) with default parameters and only reads with
mapping quality greater than 3 were used for the downstream analysis. Genome Analysis Toolkit (GATK,
version 3.3-0-g37228af) (5) was used to call single nucleotide polymorphisms (SNPs) and indels with the
parameter “--stand_call_conf 30 -stand_emit_conf 10” after realignment at the indel loci as recommended.
All SNPs were filtered based on mapping and SNP quality and sequence depth using the parameter “--
clusterWindowSize 10 MQO >= 4 & & ( MQO / (1.0 * DP)) > 0.1) QUAL < 30.0 QD < 5.0”, which
resulted in 16,904 polymorphic markers. For linkage map construction, we further removed all markers
that were missing in more than 30% of individuals or showed segregation distortions (*P* < 0.001, *χ*² test).
The final dataset used for the linkage group construction contained 7989 markers. The linkage map was
constructed using ASmap (6) following recommended workflow. In brief, all markers that were typical
for more than 70% of individuals were used to construct the backbone of the linkage map using
parameters: dist.fun = “kosambi”, p.value = 1e-12. These parameters were selected based on several
rounds of manual optimizations. Then, markers that were typical of less than 70% of all individuals were
pushed back to the map based on their similarity with markers included in the backbone map. The final
linkage map consists of 12 linkage groups with 2,906 cM.
1.4 Anchor scaffolds to pseudomolecules

To anchor scaffolds to pseudomolecules, we first used chromonomer (http://catchenlab.life.illinois.edu/chromonomer/) to identify conflicts between genome assembly and linkage maps and remove inconsistent markers. In total, 219 scaffolds were identified as containing potential assembly errors and 188 were split at the largest gap between two markers that mark the boundaries of the two positions. The remaining 31 scaffolds that contained too many potential errors were excluded from anchoring to pseudomolecules but included as scaffolds in the final genome assembly. The final step of anchoring scaffolds to pseudomolecules was performed using ALLMAPS(7). In total, 2,132 scaffolds representing 825.8Mb (34.9%) of the final genome assembly and 13,078 (38.9%) of total predicted genes were anchored to 12 pseudomolecules (Fig. S2).

The final assembly consists 37,194 scaffolds, which represents 92% (2.37 Gb) of the estimated genome size. The N50 scaffold and contig sizes equal to 524.5kb and 64.2kb, respectively, and 50% of the genome was represented in 420 longest scaffolds (L50). The detailed step-by-step workflow of the genome assembly procedures is shown in Fig. S3.

1.5 Quality assessment and validation of assembly

We assessed the completeness and quality of the final N. attenuata assembly using four different methods. First, we mapped 80,044 N. attenuata EST sequences that were assembled in a previous study (with a length greater than 300bp) to the genome using gmap(8). The overall mapping rate was ~97.1% (identity greater than 95% and coverage greater than 80% of each transcript). In addition, we also mapped 1207 million paired-end reads to the genome using TopHat2(9). Overall, 96.5% of them could be mapped to the genome and 94.2% were properly paired. Furthermore, among the 45,816 N. benthamiana EST sequences from NCBI (length greater than 300 bp), 88.7% of them could be mapped to the final N. attenuata assembly (greater than 85% identity and 60% coverage). Second, mapping 224 million (a subset) paired-end 100 bp Illumina reads (1kb library) back to the genome using Bowtie2 (-I 500 -X 2000) showed that 99.94% of reads can be mapped to the genome and 97.7% were properly paired. Third, we
used CEGMA based on 248 ultra-conserved CEGs to further estimate the completeness of the assembly.

In total, 208 (83.9%) were found in full length and 243 (98.0%) were found to be partial or full-length in our final assembly. The overall completeness of the final *N. attenuata* assembly estimated based on CEGMA is slightly less than the tomato genome, but similar to the potato genome (Table S2). Fourth, re-sequencing of ~2kb the upstream regions of 26 candidate genes using Sanger sequencer confirmed that 96.2% (25 out of 26) of the assembly were correct. For these 25 correctly assembled genes, the similarity between sequences obtained from Sanger sequencer and WGS assembly is more than 99.86%. Most of the mismatches (53 out of 72 bps) were due to an additional AT-rich fragment was miss-assembled to the promoter region of NIATv7_g09977. Together, these data suggest that the overall quality of this assembly is high.

2. Genome annotation

2.1 Annotation of repeats and LTR insertion time estimation in *Nicotiana*

The summary of repeat content is shown in Table S4. The abundance of MITEs is shown in Supplementary Dataset 1. Because the expansion of LTRs contributed to the rapid increase of genome size in *Nicotiana*, we further analyzed the insertion time of this subgroup. All annotated LTR/gypsy and/or LTR/copia TE sequences with a mapping score greater than 250 and a size greater than 100 bp were extracted from the RepeatMasker output. Only TE families that contained more than 200 copies were retained for downstream analysis. For each of the extracted LTR sequences, BLASTN was used to find the other LTR sequences with the highest similarity score with parameters: -evalue 1e-10 -task dc-megablast. The pairwise aligned fragments that had a length greater than 200 bp and a bit score greater than 200 were used for estimating substitution rates. The identical reciprocal best blast hits were only counted once. A substitution rate was then calculated using baseml from the PAML (4.7) package with the “REV” molecular evolution model. Our LTR expansion dating approach is different from the method described by SanMiguel et al(10), which compares the sequence divergence between two LTR regions of
complete retrotransposons. However, in the *N. attenuata* genome like in many other plant genomes, the
complete LTR set only contributes to a small proportion of the total amount of retrotransposons,
especially since many old retrotransposons are rapidly disrupted by new insertions and thus cannot be
detected as complete LTRs. This may result in a biased picture of the overall insertion history of
retrotransposons in the genome. Our method directly calculates the divergence between most recently
duplicated repeat sequences and thus avoids the bias introduced from annotating complete LTRs.

2.2 Analyzing DTT-NIC1 insertions in *Nicotiana* genomes

To analyze the effect of DTT-NIC1 insertions on gene expression divergences between
duplicated genes, we first calculated the gene expression and tissue specificity divergences between
duplicated gene pairs in *N. attenuata* using 21 different RNA-seq libraries (see details below on
expression analysis and gene duplication identification). To reduce redundancy and false positives, we
performed the analysis based on following protocol: 1) we analyzed the gene duplication history of all *N.
* attenuata genes based on the constructed phylogenetic trees of each homolog group; 2) only the
reciprocal most recently duplicated paralog pairs were identified and used for the downstream analysis; 3)
only the gene pairs that resulted from duplications that enjoyed an approximate Bayesian support of
greater than 0.9 were retained in the analysis; 4) all gene pairs were assigned into two groups: A: one
member of the gene pair with at least one DTT-NIC1 insertion (within the 1kb upstream of 5’ region) and
B: neither one of gene pair with DTT-NIC1 insertions; 5) gene expression divergence at both expression
and tissue specificity levels were then compared between the two groups. The tissue-specificity score for
each gene/transcript using the τ index (11) based on a subset of the RNA-seq libraries.

2.3 Annotation of miRNA, tRNA and rRNA

Small RNA sequencing was performed to capture all small RNAs (smRNAs), especially the
miRNAs that are expressed during day and night in the wild-type *N. attenuata*. Three replicates were used;
clean reads were generated from the raw smRNA reads after removing the adaptor sequences and
removing the low quality reads—such as reads with unidentified nucleotides (N) or reads with any single
nucleotide stretch > 5 nucleotides. Clean reads, > 15 nucleotides, were filtered for further analysis. Next, all the clean reads were aligned against Rfam and the reads mapped to tRNA, rRNA and snoRNA were removed. Remaining reads in all the replicates were merged for each time point, and reads that were expressed in at least two of three replicates were retained for further analysis. These reads are referred to as “mappable miRNA reads.” These reads were aligned to the N. attenuata genome using Bowtie with maximum two mismatches and five reported genomic location alignments. Sequences that did not match the genome were discarded. Further, aligned reads were mapped against the N. attenuata transcriptome by Bowtie with no reverse complement mapping parameter, and those aligned to the transcriptome were removed from the bin of miRNA mappable reads. Conserved miRNAs were identified by comparing the N. attenuata mappable reads against the known plant mature miRNAs and their precursors in the miRBase database (www.mirbase.org). Sequences with perfect matches to the known sequences were regarded as bona-fide conserved miRNAs and were subjected to precursor prediction in the Nicotiana genome. A flanking sequence of 200 bases was extracted from the mapped genomic locations for each read and RNAfold was used to predict precursor stem-loop structure. The miRCheck algorithm was used to compare all the potential miRNAs and precursors against a set of secondary structure constraints derived from known plant miRNA precursors.

A total of 131 miRNA reads corresponding to 83 bona-fide conserved miRNAs were identified that were expressed during day and night harvests of N. attenuata leaves. These miRNAs were obtained from 158 genomic locations in Nicotiana genome and were conserved in 34 other plant species (such as Arabidopsis lyrata, Arabidopsis thaliana, Vitis vinifera, and Zea mays etc.) as shown in Supplementary Dataset 2. Of these 83 miRNAs (131 miRNA reads), twelve miRNAs (miR160c-3p, miR171b-3p, miR398a-3p, miR426, miR1429-5p, miR5069, miR5497, miR6020b, miR6021, miR6206, miR7711-5p, and miR7744-5p) were expressed only during day time. On the other hand, 14 miRNAs (miR160-5p, miR408-5p, miR2609a, miR4379, miR5015, miR5042-3p, miR5255, miR5303c, miR5635a, miR5741a, miR6444, miR7750-5p, miR8143 and miR8764) were expressed only during the night. Fifty seven miRNAs (corresponding to 88 miRNA reads) were expressed during both day and night, of which 12
miRNAs (miR172c, miR172j, miR5303, miR5303a, miR6149a, miR6161a, miR6161b, miR6161c, miR6161d, miR6164a, miR6164b and miR7997c) were expressed with different mature sequences at these times. Differences in sequences for a miRNA indicate that different isomiRs were expressed between the day and night.

tRNAs were first annotated using tRNAscan-SE (v 1.3.1)(12) with default parameters. All predicted pseudo tRNAs and tRNAs that showed high similarity to chloroplast and mitochondrial genomes (e-value < 1e-10, identify > 95) were removed. In total, 1052 tRNAs that decode standard 20 AA and four selenocysteine tRNAs were predicted. rRNAs were annotated using RNAmmer (v1.2)(13) with parameter “-S euk -m tsu,lsu,ssu”. In total, 799 rRNA sequences were predicted.

2.4 Gene and transcript expression analysis

In total, 21 RNA-seq library from different tissues and same tissue with different treatments were sequenced (Table S5). Expression profile of all predicted gene models is provided in Supplementary Dataset 3. For transcript expression, only the transcripts with TPM greater than one in at least one sample were considered as expressed. In total, 25,506 genes and 73,624 transcripts were expressed, respectively (Supplementary Dataset 4). Among all sequenced tissues, roots and pollen tubes expressed the highest and lowest number of both genes and transcripts (Fig. S15), respectively. We further annotated the repeat content of all expressed transcripts using RepeatMasker (open-4.0.5) with repeat library annotated from N. attenuata as described previously. Among these 73,624 expressed transcripts, 17,463 of them (~24%) were found containing repeats (Supplementary Dataset 4), suggesting that a large number of repeats of N. attenuata are expressed.

The floral gene expression was calculated based on RNA-seq library from open flowers (OFL), and for tissues that contain more than one library (with different treatments), we used the average expression values among all libraries to represent the expression level in the respective tissues. Tissue-specific genes/transcripts were considered as t index >= 0.95. Among all sequenced tissues, roots have the highest number of tissue-specific genes and transcripts (Fig. S16). Interestingly, although pollen tubes
expressed fewer genes (4201) than other tissues, they harbored the largest proportion of tissue-specific
genes (10%).

We annotated all of the alternative splicing (AS) events using JunecBASE (v0.6) (14) based on the
splicing junctions (SJ's) information from the BAM files produced by Tophat2 (9). To reduce false
positives that likely result from non-specific or erroneous alignments, all original junctions were filtered
based on overhangs greater than 13 bp in length, as shown in our previous study (15). The overall pattern
of AS detected among different tissues is consistent with our previous study based on roots and leaves of
N. attenuata, which showed that intron retention (IR) and exon skipping (ES) are the most and least
abundant AS events, respectively (Fig. S17).

2.5 Expression of transposable elements

The expression of TE was analyzed using two different datasets. Among different tissues, RNA-
seq data were used. We first remapped all clean reads to the N. attenuata genome using tophat2 with the
parameters mentioned above, except for the retention of 150 multiple mapped reads (~g 150). Then, the
mapped bam files were used to estimate expression of the different TE families using TEtranscripts (16).
Among all measured tissues, we found that TE expression was highest in pollen tube (Fig. S18).

For M. sexta herbivory-induced TE expression in leaves, we used a previously published time-
course microarray dataset (17). All probes were mapped to the genome using blastn. Only probes that had
a perfect match to the genome were considered. Overlaps between positions of mapped probe and
repeats were identified using bedtools (18) based on repeatmasker output. All probes of which 100%
could be located within the repeat region and did not map to any annotated genes were used to analyze
expression of the M. sexta herbivory-induced expression of TEs. The microarray data were first
normalized using quantile normalization, and differential gene expression was analyzed using the limma
R package (19). Probes that showed a false discovery rate (FDR) adjusted $P$ value lower than 0.05 and a
fold change greater than 2 were considered as induced by M. sexta OS. Differentially expressed probes
that were annotated as repeats are reported in Supplementary Dataset 5. Similar to protein coding genes,
most of TEs showed the highest induction at 1 h after elicitation (Fig. S19), except LTR/copia, which
were induced the most at 13 h after elicitation.

2.6 Functional annotation of protein coding genes

The functions and gene ontology (GO) of all predicted protein coding genes were annotated
using BLAST2GO(20). All protein coding sequences were first compared to the plant reference sequence
database (downloaded in February 2016) using BLAST with e-value cutoff 1e-10. The BLAST results
were imported to BLAST2GO and the functions and GO terms of each gene were annotated using the
default settings. In total, 59.3% of genes were assigned to at least one GO term. The enzyme commission
(EC) number was also assigned to the predicted genes by comparing to the KEGG pathway databases.
The pathways that contained only one mapped enzyme were removed. Overall, 5,370 genes (16.1 %)
were assigned to 942 EC codes that were involved in 144 pathways. The top three pathways that contain
the highest number of annotated genes are starch and sucrose metabolism, purine metabolism and
phenylalanine metabolism. Furthermore, the KEGG orthologs (KO) were annotated using kobas 2.0 (21),
with e-value cutoff set as 1e-10. In total, 10,746 genes were assigned to KO terms. The information of
mapped KO information for each gene is reported in the Supplementary Dataset 6.

2.7 Functional annotation of protein domains

To identify the functional domains of protein-coding genes, INTERPROSCAN (v. 5.16-55.0)(22)
was used to scan protein sequences against the protein signatures from InterPro database (v. 55.0,
downloaded in February, 2016). The InterPro database integrates protein families, Pfam domains and
functional sites from different databases. The following databases from InterPro were used: Coils (v.
2.2.1), CATH-Gene3D (v. 3.5.0), HAMAP (v. 201511.02), Panther (v. 10.0), Pfam (v. 28.0), PIRSF (v.
3.01), PRINTS (v. 42.0), ProDom (v. 2006.1), PROSITE (v. 20.113), SMART (v. 6.2), SUPERFAMILY
(v. 1.75) and TIGRFAM (v. 15.0). In total, 38,810 protein domains were identified, consisting of 3,925
unique Pfams. For predicted genes in N. attenuata genome, 71.1% (23,783 out of 33,449) of them contain
at least one predicted protein domain. The top 20 Pfam and SUPERFAMILY domains are shown in Fig.
S20 and S21. The top 20 most frequent domains account for 22.7% (8,780 out of 38,729) and 37.6%
(10,950 out of 29,091) of all predicted Pfam domains and SUPERFAMILIES respectively. In comparison
to tomato(23), most of the top 20 SUPERFAMILIES are similar except three, which are ribonuclease H-
like (SSF53098), DNA/RNA polymerases (SSF56672) and trans-glycosidases (SSF51445). At domain
levels, the top 20 list in N. attenuata and tomato differed in seven domains, including domains of reverse
transcriptase-like (PF13456), PPR repeat (PF12854) and a domain of unknown function (PF14111).

2.8 Annotation of transcription factor and protein kinase genes

The transcription factor and protein kinase-containing genes were identified based on the
identified domain in each gene according to rules described in Pérez-Rodriguez et. al(24) using the iTAK
tool (http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi). In total, 2,498 and 1,071 genes were annotated
as transcription factor and protein kinase genes respectively (Supplementary Dataset 7). Among all
transcription factors, MYB, AP2 and bHLH were the three largest families.

3. Evolutionary genomics analyses

3.1 Homolog group identification

In total, we identified 23,340 HGs (with at least two homolog sequences) from the 11 plant
genomes (Table S1). The average gene family size within each species varied from 1.8 in C. sativus to 2.8
in P. trichocarpa. Among these identified HGs, 4,328 contained at least one gene from each of all 11
dicots species, thus representing the core dicot HGs. In N. attenuata, 13,632 HGs containing 30,513
protein-coding genes were identified, of which 89.7% (12,231) contained orthologs from both Nicotiana
genomes sequenced in this study. In addition, 8.8% (2,936 of 33,449) of N. attenuata genes were assigned
as orphan genes, since they did not cluster with any other sequences (Fig. S22).
3.2 Detection of gene duplication events in each HG

We first constructed the phylogenetic tree for each HG. For this, we aligned all coding sequences for each HG using MUSCLE (v.3.8.31) (25), based on translated protein sequences from TranslatorX (v.1.1) (26). For all aligned sequences, all non-informational sites (gaps in more than 20% of sequences) were removed using trimAl (v.1.4) (27). Then, for each HG, PhyML (v. 20140206) (28) was used to construct the gene tree with the best nucleotide substitution model, estimated based on jModeltest2 (v.2.1.10) (29) with the following parameters: -f -i -g 4 -s 3 -AIC -a. The support for each branch was calculated using the approximate Bayesian method.

The duplication events within each HG were predicted based on the constructed gene trees using a tree reconciliation algorithm which compares the structure of a species’ tree and gene tree to infer the duplication events(30). This approach allowed us to predict the history of gene duplication events on each branch of the species tree. To reduce false positives, we only considered tree structures with approximate Bayes branch supports greater than 0.9 for the downstream analysis. In addition, we also characterized the gene duplication events based on their genomic location. Tandem duplicated gene pairs were identified as gene pairs from the same HG if located within a distance of 100 kb or four genes from each other.

Using the tree reconciliation algorithm mentioned above, we identified 81,859 duplication events on the branches of 11 species tree. Among all these duplicated events, genes from *N. attenuata* were involved in 27.0% of them (22,054 out of 81,859). The species/genera that were known to have experienced whole genome duplication (WGD) or triplication (WGT), such as *Arabidopsis, Populus, Mimulus*, showed much higher numbers of duplication events (Fig. S4, and Supplementary Dataset 8).

Furthermore, a large number of gene duplication events were found on the branch of Solanaceae, indicating the shared WGT (23) between the *Nicotiana* and *Solanum* genera (Fig. S4). Although the cultivated potato is an autotetraploid, the genome sequences used for our analysis represent a diploid species (2n=2x=24), which were obtained through the combinations of sequencing a homozygous doubled-monoploid derived from classical tissue culture and a heterozygous diploid breeding line (31).
Therefore the additional polyploidization likely did not affect the detection of gene duplication events at the branch of Solanaceae and *Nicotiana*.

Additionally, we inferred the type of duplication to be tandem duplication or genome-wide duplication by using gene distance and syntenic block information of the genome, respectively. All duplicated genes gene pairs that are separated less than 10 genes and were less than 150 kbp apart in the genome of any species of tomato, potato or *N. attenuata* were considered as derived from tandem duplications. In total, we found 535 duplications that occurred on the Solanaceae branch (7.8%) which were likely to be tandem. We further used syntenic blocks estimated in tomato and potato that were predicted by the Plant Genome Duplication Database (PGDD, [http://chibba.pgml.uga.edu/duplication/index/downloads](http://chibba.pgml.uga.edu/duplication/index/downloads)) to infer duplications that resulted from genome-wide duplications. All duplication events occurred on the Solanaceae branch and can be found in a syntenic block (with e-value < 0.01 and 0.3<Ks<1.7) were considered as derived from whole genome duplications.

One example of duplications via WGT is the evolution of *threonine deaminase* (TD) family, which has four copies in both *Nicotiana* and *Solanum* resulting from one round of local duplication followed by a genome-wide duplication (Fig. S5). TD is a Solanaceae specific anti-herbivore defense gene that encodes a pyridoxal phosphate-dependent enzyme that dehydrates threonine to α-ketobutyrate and ammonia, as the committed step in the biosynthesis of isoleucine (Ile) in plants. After one tandem duplication followed by WGT, the duplicated gene copy (*TD2.2*), likely through several adaptive substitutions (32), evolved as an anti-digestive defense in tomato that degrades threonine in the guts of attacking larvae (33). Interesting, in *N. attenuata* *TD2.2* likely retained its biosynthetic function to supply the Ile essential for jasmonate-mediated defense signaling (34). Consistent with a functional specialization and in contrast to other WGT-derived *TD* copies, *N. attenuata* *TD2.2* exhibits a unique spatial expression pattern and is specifically induced by herbivore attack (Fig. S5).
3.3 Confirmation of the whole-genome triplication in Nicotiana

To further support the conclusion that *Nicotiana* spp. share the WGT event previously identified in *Solanum* (23), we performed molecular evolution and phylogenomic analyses. For the molecular dating, the synonymous substitution rates (Ks) for the gene pairs in all homologous groups (with less than 5 genes per species) were calculated based on YN model using KaKs_Calculator (v.1.2) (35). To estimate the divergence time, the formula $T = \text{Ks}/(2\times\text{mutational rate})$ was used, where the mutational rate was set to $7.1 \times 10^9$ substitutions/site/million years (MYA)(36). The within-species 4dTV distribution in *N. attenuata* and tomato showed a similar peak between 0.2 to 0.25, indicating that a genome-wide duplication event had occurred at a similar time in both species (Fig. S6). Given the previously identified WGT in tomato, it is likely that *Nicotiana* shares this genome-wide duplication event with tomato. This inference is also consistent with the more ancient estimated age of this WGT (71 ±19.4 MYA)(23) compared to the recent divergence between *Nicotiana* and *Solanum* (24.4 MYA to 28.6 MYA).

To further validate the shared WGT event between *Nicotiana* and *Solanum*, we analyzed a subset of genes that underwent a triplication in either *Solanum* or *Nicotiana*. The genes that underwent a triplication event in *Solanum* were identified based on the fact that the tree structure consists of one outgroup node (at least one of genes in *Arabidopsis*, *Populus*, *Cucumber* or *Vitis*) and two duplication events that were shared between tomato and potato with no gene loss in either tomato or potato. Similarly, the genes that underwent a triplication event in *Nicotiana* were identified based on the fact that the tree structure consists of one outgroup node and two duplication events that were shared between *N. attenuata* and *N. obtusifolia* with no gene loss in either species. In order to reduce the false positives, we only considered the tree nodes that have approximate Bayes branch supports of greater than 0.9. We then analyzed the number of these triplication events that were shared between *Solanum* and *Nicotiana* based on the complete tree structure. In total, we identified 229 and 436 triplication events in *Solanum* and *Nicotiana*, respectively. In both the *Solanum* and *Nicotiana* datasets, the majority (89.2% and 79.9%) respectively) of the triplication events were shared with the other genus. These results are consistent with the fact that the WGT event found in tomato is indeed shared between *Solanum* and *Nicotiana*.
Furthermore, using MCSCAN, we also found 22 duplication blocks, each of which contains at least 20 genes among assembled pseudo-chromosomes (Fig. S7). In addition, we also compared the identified triplication events in *Solanum* and *Nicotiana* with those of genes in *Mimusus*. Our results showed that a majority of the triplication events (98.7% and 98.4% in *Solanum* and *Nicotiana*, respectively) were not shared with *Mimusus*, indicating that gene duplication events in *Mimusus* are independent of the WGT found in Solanaceae.

### 3.4 Estimation of species divergence times

We calculated the divergence times of four Solanaceae spp. (*N. attenuata*, *N. obtusifolia*, *S. lycopersicum* and *S. tuberosum*) with *V. vinifera* as the outgroup using a Bayesian approach. In brief, we first identified 1,622 one-to-one orthologs using BLAST reciprocal best-hits (RBH) algorithm. Then each orthologous group was aligned using the protein coding sequences with MUSCLE (v. 3.8.31)(25) based on translated protein sequences from TranslatorX (v.1.1)(26). These alignments were concatenated to one super-alignment and all ambiguously aligned regions were removed using trimAL (v. 1.4) (27) with parameter: -gt 0.8. The final alignment that contained ~2.1Mb nucleotide sites was then used for species divergence time estimation. The nucleotide substitution parameters were first estimated with the HKY85 model using baseml from the PAML package (v4.8)(37). The branch-length and the corresponding variance-covariance matrix were then estimated using estbranches. The results from estbranches were used to predict divergence times with the MultiDivTime program(38). The MultiDivTime analysis was performed according to the manual with the following parameters: the root-to-tip mean was set to 119.5 MYA with a standard deviation of 10 MYA based on the divergence time estimated by Guyot et. al(39); the evolutionary rate of the root was set to 0.007631799 substitutions per nucleotide site per MYA calculated based on the results of baseml; and a burn-in time of 10,000,000 generations was used. MCMC chains were sampled every 10,000 generations until 100,000 samples were taken and the calibration points of 7.3 MYA and 23.7 MYA for the divergence time of tomato-potato and tomato-*Nicotiana*, respectively(40), were used.
The analysis revealed that *N. attenuata* and *N. obtusifolia* diverged about 12.5 MYA ago (95% confidence interval: 10.1 MYA to 14.7 MYA), which is about five million years earlier than the divergence time between potato and tomato (7.1 MYA, range from 6.4 MYA to 8.2 MYA). The estimated divergence time between the *Nicotiana* and *Solanum* lineages was estimated to be of 27.1 MYA (95% confidence interval: 24.4 MYA to 28.6 MYA) and the divergence time of *V. vinifera* and solanaceous species is around 116.70 MYA (95% confidence interval: 100.9 MYA to 133.6 MYA).

### 3.5 Identification of lineage-specific gene family expansion

We analyzed Solanaceae and *Nicotiana* lineage-specific HG expansions using the previously identified gene duplication events. For this, we performed a Fisher’s exact test to identify gene families that exhibit significantly more duplications in comparison to the genome-wide pattern. Because the total number of gene families was large, we reduced the number of tests and false positives for a given branch by calculating a Z-score for each HG and only considering those HGs that had Z-scores > 1.96 for Fisher’s exact test. Multiple testing corrections were then performed based on these P-values using the false discovery rate (FDR < 0.1) method. It should be noted that this is a conservative approach, as many small gene families cannot be detected by such stringent statistical tests. Hence, the gene families identified by this analysis as significantly expanding are likely true positives.

For the Solanaceae branch, we found 1,596 HGs with a Z-score greater than 1.96. Among them, four HGs experienced significantly more duplications than the genome-wide pattern based on Fisher’s exact test. One of these belongs to the *S-locus F-box protein (SLF)* family, of which 10 duplications were identified on the branch of Solanaceae. Consistently, a recent study also found increased number of SLF genes in *Petunia*, another Solanaceae species(41). The phylogenetic tree combining SLF from *Petunia*, *Solanum* and *Nicotiana* revealed that this particular SLF subfamily experienced frequent duplication and gene loss in different lineages (Fig. S23). The SLF gene family is known to be involved in pollen recognition and self-compatibility processes, and the expansion of this family might have contributed to the observed diversity of the mating systems in the Solanaceae family(42). Another HG that significantly
expanded in the Solanaceae is corresponding to the Zeatin O-glucosyltransferase-like (ZOG) gene family, of which we detected 25 duplication events (Fig. S24). Genes from the ZOG family are involved in regulating cytokinin levels a process critical in tuning the signaling mediated by this hormonal pathway to biotic and abiotic stresses. Expansion of this gene family in the Solanaceae branch likely reflects physiological adaptations to the diverse habitats colonized by these species.

Analyzing gene duplications within the *Nicotiana* branch showed that 256 HGs have a Z-score greater than 1.96, and 58 HGs experienced significantly more duplications, based on Fisher’s exact test, than seen from the genome-wide pattern (Supplementary Dataset 9). Among these significantly expanded gene families, an NBS-LRR type disease resistance gene family specifically duplicated three times in *Nicotiana* but not in other branches of the Solanaceae. Genomic location of these duplicated genes shows that all four genes are located in a gene cluster within one scaffold (Fig. S25). Further expression analyses revealed that all four genes are highly expressed in roots, suggesting that these genes might be involved in plant-pathogen interactions in *Nicotiana* roots.

Another gene family that significantly expanded in *Nicotiana* species is that of the Purine uptake permeases (Fig. S26). Genes from this family are known as plasma membrane-localized transporters. More specifically, in tobacco, one member of this family, *Nicotine Uptake Permease 1* (NUP1), has demonstrated functions in regulating nicotine localization via its transporter activity. Furthermore, *NUP1* has also been shown to act as a transcriptional regulator of the key transcription factor *ERF189* in the nicotine biosynthesis pathway, although details on the underlying mechanism are lacking. The expression profile of *NUP1* in *N. attenuata* highlights that this gene is not only expressed in roots, but also shows high levels of expression in floral tissues (Fig. S26), which indicates that *NUP1* might be involved in the allocation of nicotine to flowers of *N. attenuata*, where it serves as a deterrent for pollinators and increases out-crossing rates. In addition, several other members of this family are specifically induced in the leaves and stem by simulated herbivory, indicating that these genes in *N. attenuata* might also be involved in allocating nicotine to particular parts of the vegetative tissues as an anti-herbivore toxin.
4. Evolution of nicotine biosynthesis

4.1 Identification and reconstruction of the evolution history of nicotine biosynthesis genes

We identified nicotine biosynthesis genes and their ancestral copies based on sequence homology using blastp and manual curations (Table S6). Gene evolutionary history was inferred using the combination of phylogenetic and synteny analysis when possible. However, due to lack of genomic information from the closely related genus of *Nicotiana*, it remains unclear whether the *Nicotiana*-lineage specific duplications, such as the NAD pathway genes and neofunctionalization of *BBLs* occurred in the ancestors of all genera, in which trace levels of nicotine have been controversially reported, such as *Crenidium*, *Cyphanthera* and *Duboisia* (48).

The putrescine biosynthetic pathway for nicotine biosynthesis could have two routes: 1) synthesized by ODC-mediated decarboxylation of ornithine or 2) synthesized by ADC-mediated decarboxylation of arginine. Recent studies that individually silenced ODC and ADC suggest that the putrescine for nicotine biosynthesis in *Nicotiana* is likely through the former route (49, 50). Consistently, while phylogenetic analysis showed that two ADC copies are present in *Nicotiana* genomes likely through the WGT, none of them showed root specific expression pattern.

4.2 Validation of the promoter sequences of nicotine biosynthesis genes and their ancestral copies

We validated the 2kb promoter sequences of 26 genes from *N. attenuata*, which included all nicotine biosynthesis genes and several of their ancestral copies using Sanger sequencing. Primers were first designed based on the assembled genome sequences, and direct PCRs were performed to amplify the target fragments (Table S7). For genes that gave multiple bands or no PCR products, nested PCRs were performed. The amplified fragments with the expected size were cut from gels and used for either direct sequencing or sequencing after cloning into the pJET vector. All Sanger sequences were manually inspected and compared to the genome assembly. Among all tested genes, only the promoter region of one gene, NIATv7_g05934, was found miss-assembled (PCR products do not match the genome.
assembly). For all other correctly assembled genes, 99.86% identity was found between Sanger sequencing and the assembled genome.

**4.3 Prediction of TE-derived putative miRNA target sites into regulatory region of nicotine biosynthesis genes**

Insertions of TE are known to introduce putative target sites of miRNAs. To examine whether the observed TE insertions within the 2kb upstream region of nicotine biosynthesis genes and of their ancestral copies had introduced candidate miRNA targeting sites, we performed *in silico* miRNA target site predictions using the method described in Pandey et al (51). Briefly, all the candidate promoter sequences were first checked for miRNA seed-pairs using custom written Perl script. The promoter sequences from 3’ end were used for "Watson-Crick" complementarity matching against 5’ ends of miRNAs after generating 7- to 13-nt seeds starting from first or second nucleotide position at 5’ end of the miRNAs. The matches were extended by allowing mismatches after the seed match or 9th nucleotide in the miRNAs.

In total, 23 and 17 putative miRNA target sites were predicted among the 2kb upstream regions of 10 genes that are involved in nicotine biosynthesis and of their ancestral copies or non-root specific expressed genes (10), respectively (Table S8). Among all nicotine biosynthesis genes, five predicted miRNA target sites that were detected from 4 genes (*BBL2.2, PMT1.2, ODC2* and *QS*) are overlapped with TE insertions.

**4.4 Evolution of root-specific expression of nicotine biosynthesis genes**

One of the important features associated with efficient nicotine production is the root-specific expression of nicotine biosynthesis genes. To obtain more evolutionary insights into the process of root-specify expression acquisition by these genes, we explored the expression pattern of orthologous genes of the nicotine biosynthetic pathway in tomato and potato. The results show that although absolute expression levels of these orthologues are consistently low in the roots of these two species, some genes
exhibit a certain degree of root-specific expression in these Solanum species, such as PMT1 and ODC2 from the duplicated polyamine pathway as well as the homologs of BBLs and A622. These observations indicate that the evolution of a preferential root expression for these genes may have taken place before the divergence between Nicotiana and Solanum (Fig. S27). This, however, does not exclude the likely possibility that the neofunctionalization of BBL and A622 occurred after the divergence between Nicotiana and Solanum. In contrast, the root-specific expression of the NAD pathway genes, such as AO2, QPT2 and QS likely evolved specifically in Nicotiana. A further comparative analysis of the expression profile of ERF189, the key transcription factor that regulates most of the nicotine biosynthesis genes, showed that its root-specific expression likely evolved in Nicotiana, as its orthologues are more uniformly expressed across tissues in tomato and potato (Fig. 4).

From a mechanistic point of view, there are at least two possibilities: 1), while gaining additional GCC/G-box motifs increased their expression in roots, nicotine genes may have lost the transcription factor binding sites that allow them to be expressed in other tissues, and 2) specific smRNAs may have contributed either directly or through interactions with endogenous target mimicry (52) the suppression of the expression of nicotine genes in non-root tissues. We speculate that both mechanisms might have contributed to root expression specialization. Future molecular studies that specifically manipulate the promoter sequences of nicotine biosynthesis genes and targeted smRNAs will be required to understand this process.

4.5 Data access

To provide free access to our genome data, we established a webservice (Nicotiana attenuata Data Hub: http://nadh.ice.mpg.de/) (53), which allows data visualization of gene families and co-expressed genes, as well as data download of the annotated gene models of the two Nicotiana genomes. The Whole Genome Shotgun projects of N. attenuata and N. obtusifolia have been deposited at DDBJ/ENA/GenBank under the accession MJEQ00000000 and MJEQ00000000, respectively. All short reads and PacBio reads used in this study were deposited in NCBI under BioProject PRJNA317743 (RNA-seq reads of N.
attenuata), PRJNA316810 (short gun reads of N. attenuata), PRJNA317654 (WGS PacBio long reads of
N. attenuata), PRJNA316803 (RNA-seq reads and assembly of N. obtusifolia), and PRJNA316794 (short
reads of N. obtusifolia). The assembled genome sequences of N. attenuata and annotation information of
all protein coding genes are available from the CoGe platform, which provides genome-browser view and
downstream comparative genomic analysis, and Sol Genomics network (https://solgenomics.net/). There
are 129 gene models that are deposited in CoGe and SGN but were not submitted to NCBI because they
contain introns smaller than 10bp, which do not fulfill the criteria for NCBI submission. However these
genes models are likely to be correct based on their sequence conservation. Therefore, they were retained
and made them publically available via SGN and CoGe. Including or excluding these gene models does
not affect any conclusion of the study. The python scripts used for phylogenomic analysis are deposited
on github (https://github.com/thobrock/Nicotiana-attenuata-genome-project/).
5. Supplemental Tables

Table S1. Genomes of 11 plant species used for comparative genomic analysis.

<table>
<thead>
<tr>
<th>Species</th>
<th>Version</th>
<th># of gene models</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>N. attenuata</em></td>
<td>r2.0</td>
<td>33,449</td>
<td><a href="http://mutb.ice.mpg.de">http://mutb.ice.mpg.de</a></td>
<td>This study</td>
</tr>
<tr>
<td><em>N. obtusifolia</em></td>
<td>v1.0</td>
<td>27,911</td>
<td><a href="http://mutb.ice.mpg.de">http://mutb.ice.mpg.de</a></td>
<td>This study</td>
</tr>
<tr>
<td><em>C. annuum</em></td>
<td>v2.0</td>
<td>35,336</td>
<td><a href="http://phytozone.jgi.doe.gov/cucurbit">http://phytozone.jgi.doe.gov/cucurbit</a></td>
<td>Kim et al. 2014 (55)</td>
</tr>
<tr>
<td><em>C. sativus</em></td>
<td>v1.0</td>
<td>21,503</td>
<td><a href="http://phytozone.jgi.doe.gov/cucurbit">http://phytozone.jgi.doe.gov/cucurbit</a></td>
<td>Huang et al. 2009 (56)</td>
</tr>
<tr>
<td><em>M. truncatula</em></td>
<td>v2.0</td>
<td>28,140</td>
<td><a href="http://phytozone.jgi.doe.gov/trichodes">http://phytozone.jgi.doe.gov/trichodes</a></td>
<td>Hellsten et al. 2013 (57)</td>
</tr>
<tr>
<td><em>P. trichocarpa</em></td>
<td>v3.0</td>
<td>41,335</td>
<td><a href="http://phytozone.jgi.doe.gov/poplar">http://phytozone.jgi.doe.gov/poplar</a></td>
<td>Tuskan et al. 2006 (58)</td>
</tr>
<tr>
<td><em>S. lycopersicum</em></td>
<td>ITAG2.3</td>
<td>34,727</td>
<td><a href="http://phytozone.jgi.doe.gov/tomato">http://phytozone.jgi.doe.gov/tomato</a></td>
<td>Tomato Consortium. 2012 (23)</td>
</tr>
<tr>
<td><em>S. melongena</em></td>
<td>v2.5.1</td>
<td>42,035</td>
<td>ftp://utlio.cs.cas.cz/pub/software/plant</td>
<td>Hirakawa et al. 2014 (59)</td>
</tr>
<tr>
<td><em>S. tuberosum</em></td>
<td>v3.4</td>
<td>35,119</td>
<td><a href="http://phytozone.jgi.doe.gov/potato">http://phytozone.jgi.doe.gov/potato</a></td>
<td>Xu et al. 2011 (31)</td>
</tr>
<tr>
<td><em>V. vinifera</em></td>
<td>Genoscope 12X</td>
<td>26,346</td>
<td><a href="http://phytozone.jgi.doe.gov/paper">http://phytozone.jgi.doe.gov/paper</a></td>
<td>Jaillon et al. 2007 (60)</td>
</tr>
</tbody>
</table>
Table S2. Completeness of *N. attenuata* and *N. obtusifolia*, two other published *Nicotiana* species, potato and tomato genomes (potato version 206 and tomato version 225) estimated based on 248 ultra-conserved CEGs using CEGMA.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Complete</th>
<th>Partial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Group 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. attenuata</td>
<td>83.87</td>
<td>77.27</td>
</tr>
<tr>
<td></td>
<td>646</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>2.23</td>
<td>1.86</td>
</tr>
<tr>
<td></td>
<td>63.94</td>
<td>52.94</td>
</tr>
<tr>
<td></td>
<td>243</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. obtusifolia</td>
<td>85.08</td>
<td>78.79</td>
</tr>
<tr>
<td></td>
<td>421</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>56.87</td>
<td>51.92</td>
</tr>
<tr>
<td></td>
<td>243</td>
<td>65</td>
</tr>
<tr>
<td>N. sylvestris</td>
<td>78.63</td>
<td>74.24</td>
</tr>
<tr>
<td></td>
<td>402</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>2.06</td>
<td>1.73</td>
</tr>
<tr>
<td></td>
<td>60.51</td>
<td>42.86</td>
</tr>
<tr>
<td></td>
<td>238</td>
<td>60</td>
</tr>
<tr>
<td>N. tomentospermales</td>
<td>66.13</td>
<td>77.27</td>
</tr>
<tr>
<td></td>
<td>321</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>1.96</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>56.71</td>
<td>41.38</td>
</tr>
<tr>
<td></td>
<td>188</td>
<td>57</td>
</tr>
<tr>
<td>S. tuberosa</td>
<td>83.87</td>
<td>77.27</td>
</tr>
<tr>
<td></td>
<td>302</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>1.88</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>53.85</td>
<td>41.38</td>
</tr>
<tr>
<td></td>
<td>219</td>
<td>63</td>
</tr>
<tr>
<td>S. fruticosum</td>
<td>86.49</td>
<td>82.33</td>
</tr>
<tr>
<td></td>
<td>391</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>1.82</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>47.44</td>
<td>34.55</td>
</tr>
<tr>
<td></td>
<td>219</td>
<td>63</td>
</tr>
</tbody>
</table>

54
Table S3. Summary of filtered (duplicate removal, quality clipping) WGS sequencing data used for the assemblies of *N. attenuata* and *N. obtusifolia*.

<table>
<thead>
<tr>
<th>Library type</th>
<th>PE180</th>
<th>PE250</th>
<th>PE600</th>
<th>PE950</th>
<th>MP5500</th>
<th>MP20000</th>
<th>SR454</th>
<th>PE480</th>
<th>PE1050</th>
<th>MP3500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>paired reads</td>
<td>354,327,6</td>
<td>254,673,1</td>
<td>152,657,0</td>
<td>92,178,3</td>
<td>80,509,5</td>
<td>2,005,65</td>
<td></td>
<td>478,531,3</td>
<td>222,388,2</td>
<td>108,153,56</td>
</tr>
<tr>
<td>non-paired</td>
<td>62</td>
<td>42</td>
<td>90</td>
<td>84</td>
<td>7</td>
<td>8</td>
<td>-</td>
<td>66</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Avg. length [bp]</td>
<td>103,391,4</td>
<td>30,351,12</td>
<td>68,691,92</td>
<td>77,197,1</td>
<td>75,0</td>
<td>75,0</td>
<td>644,1</td>
<td>91,3</td>
<td>77,6</td>
<td>94,6</td>
</tr>
<tr>
<td>Length sum [Gb]</td>
<td>87,5</td>
<td>97,9</td>
<td>91,7</td>
<td>88,6</td>
<td>75,0</td>
<td>75,0</td>
<td>644,1</td>
<td>91,3</td>
<td>77,6</td>
<td>94,6</td>
</tr>
<tr>
<td>Approx. seq. coverage</td>
<td>16.7</td>
<td>11.6</td>
<td>8.5</td>
<td>6.3</td>
<td>2.5</td>
<td>0.1</td>
<td>6.8</td>
<td>31.2</td>
<td>12.3</td>
<td>7.3</td>
</tr>
<tr>
<td>Approx. frag. coverage</td>
<td>13,3</td>
<td>13,3</td>
<td>19,1</td>
<td>18,2</td>
<td>92,3</td>
<td>8,4</td>
<td>6,8</td>
<td>82,0</td>
<td>83,4</td>
<td>135,2</td>
</tr>
</tbody>
</table>
### Table S4. Repeat content among six solanaceous genomes.

<table>
<thead>
<tr>
<th>TE class</th>
<th>TE type</th>
<th><em>S. tuberosum</em> (Mb)</th>
<th><em>S. lycopersicum</em> (Mb)</th>
<th><em>N. obliqua</em> (Mb)</th>
<th><em>N. benthamiana</em> (Mb)</th>
<th><em>N. sylvestris</em> (Mb)</th>
<th><em>N. attenuata</em> (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>DNA</td>
<td>10.3</td>
<td>11.9</td>
<td>36.2</td>
<td>66.5</td>
<td>69.4</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td>DNA/CMC-EnSpm</td>
<td>7.5</td>
<td>9.4</td>
<td>11.7</td>
<td>7.6</td>
<td>8.6</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>DNA/Dada</td>
<td>0.1</td>
<td>0.0</td>
<td>2.9</td>
<td>5.5</td>
<td>7.8</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>DNA/MULE-MuDR</td>
<td>3.3</td>
<td>3.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>DNA/MuL-E-MuDR</td>
<td>1.4</td>
<td>5.0</td>
<td>3.2</td>
<td>5.5</td>
<td>9.7</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>DNA/PIF-Harbinger</td>
<td>5.6</td>
<td>4.3</td>
<td>4.7</td>
<td>4.8</td>
<td>5.8</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>DNA/TcMar-Pogo</td>
<td>0.4</td>
<td>0.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>DNA/TcMar-Stowaway</td>
<td>9.2</td>
<td>4.1</td>
<td>2.7</td>
<td>3.9</td>
<td>4.0</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>DNA/hAT</td>
<td>0.1</td>
<td>-</td>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>DNA/hAT-Ac</td>
<td>7.5</td>
<td>4.1</td>
<td>9.0</td>
<td>11.1</td>
<td>12.5</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>DNA/hAT-Charlie</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>DNA/hAT-Tag1</td>
<td>1.7</td>
<td>1.2</td>
<td>1.2</td>
<td>2.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>DNA/hAT-Tip100</td>
<td>4.6</td>
<td>2.6</td>
<td>1.5</td>
<td>2.3</td>
<td>2.3</td>
<td>2.2</td>
</tr>
<tr>
<td>LINE</td>
<td>LINE</td>
<td>-</td>
<td>-</td>
<td>11.7</td>
<td>14.6</td>
<td>19.2</td>
<td>22.9</td>
</tr>
<tr>
<td></td>
<td>LINE/L1</td>
<td>10.6</td>
<td>7.9</td>
<td>18.2</td>
<td>12.0</td>
<td>18.3</td>
<td>24.9</td>
</tr>
<tr>
<td></td>
<td>LINE/L1-Tx1</td>
<td>0.1</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>LINE/L2</td>
<td>-</td>
<td>-</td>
<td>0.5</td>
<td>0.2</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>LINE/R1</td>
<td>-</td>
<td>-</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>LINE/RTE-BovB</td>
<td>5.1</td>
<td>3.3</td>
<td>8.6</td>
<td>13.0</td>
<td>15.4</td>
<td>12.4</td>
</tr>
<tr>
<td>LTR</td>
<td>LTR</td>
<td>0.1</td>
<td>-</td>
<td>83.1</td>
<td>133.7</td>
<td>129.4</td>
<td>132.5</td>
</tr>
<tr>
<td></td>
<td>LTR/Caulimovirus</td>
<td>3.7</td>
<td>2.4</td>
<td>12.5</td>
<td>1.5</td>
<td>2.4</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>LTR/Copia</td>
<td>29.1</td>
<td>48.8</td>
<td>42.6</td>
<td>61.5</td>
<td>87.1</td>
<td>131.0</td>
</tr>
<tr>
<td></td>
<td>LTR/ERV1</td>
<td>-</td>
<td>-</td>
<td>0.0</td>
<td>0.9</td>
<td>4.1</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>LTR/Gypsy</td>
<td>244.6</td>
<td>318.0</td>
<td>519.7</td>
<td>864.6</td>
<td>1171.4</td>
<td>1201.5</td>
</tr>
<tr>
<td>Others</td>
<td>RC/Heritn</td>
<td>2.4</td>
<td>0.6</td>
<td>3.1</td>
<td>0.0</td>
<td>3.0</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>Retrotransposon</td>
<td>3.8</td>
<td>3.5</td>
<td>8.1</td>
<td>3.4</td>
<td>12.6</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>SINE</td>
<td>0.7</td>
<td>0.5</td>
<td>1.3</td>
<td>13.5</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>Satellite</td>
<td>3.8</td>
<td>1.0</td>
<td>5.9</td>
<td>1.9</td>
<td>3.2</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>Unknown (Mb)</td>
<td>46.4</td>
<td>37.1</td>
<td>3.2</td>
<td>5.1</td>
<td>9.1</td>
<td>9.5</td>
</tr>
<tr>
<td>TE size (Mb)</td>
<td>482.2</td>
<td>469.2</td>
<td>792.3</td>
<td>1236.2</td>
<td>1601.6</td>
<td>1694.2</td>
<td></td>
</tr>
<tr>
<td>Genome size (Mb)</td>
<td>676.3</td>
<td>737.6</td>
<td>1223.2</td>
<td>1642.4</td>
<td>2047.5</td>
<td>2090.5</td>
<td></td>
</tr>
<tr>
<td>Percentage of TE in the genome (%)</td>
<td>59.5</td>
<td>63.6</td>
<td>64.8</td>
<td>75.3</td>
<td>78.2</td>
<td>81.0</td>
<td></td>
</tr>
</tbody>
</table>
Table S5. Information on samples used for RNA-seq.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Treatment / development stage</th>
<th>Library ID</th>
<th># raw reads</th>
<th># clean reads</th>
<th>% uniquely mapped reads</th>
<th>Read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roots</td>
<td><em>M. sexta</em> OS induced on leaves</td>
<td>NA1498ROT</td>
<td>327,772,944</td>
<td>317,843,200</td>
<td>93.77</td>
<td>50bp</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>NA1717LEC</td>
<td>61,531,550</td>
<td>16,430,982</td>
<td>93.29</td>
<td>100bp</td>
</tr>
<tr>
<td></td>
<td><em>M. sexta</em> OS locally induced</td>
<td>NA1500LET</td>
<td>328,071,888</td>
<td>317,905,162</td>
<td>91.94</td>
<td>50bp</td>
</tr>
<tr>
<td>Seeds</td>
<td>Smoke solution treated</td>
<td>NA1501SES</td>
<td>51,423,280</td>
<td>49,120,770</td>
<td>90.53</td>
<td>50bp</td>
</tr>
<tr>
<td></td>
<td>Water treated</td>
<td>NA1502SEW</td>
<td>75,944,970</td>
<td>73,525,266</td>
<td>90.31</td>
<td>50bp</td>
</tr>
<tr>
<td></td>
<td>Dry</td>
<td>NA1503SED</td>
<td>63,463,542</td>
<td>61,285,870</td>
<td>89.01</td>
<td>50bp</td>
</tr>
<tr>
<td>Stems</td>
<td><em>M. sexta</em> OS treated on leaves</td>
<td>NA1504STT</td>
<td>72,473,514</td>
<td>70,573,710</td>
<td>94.70</td>
<td>50bp</td>
</tr>
<tr>
<td>Corollas</td>
<td>Early developmental stage</td>
<td>NA1505COE</td>
<td>44,064,054</td>
<td>23,091,618</td>
<td>95.14</td>
<td>100bp</td>
</tr>
<tr>
<td></td>
<td>Late developmental stage</td>
<td>NA1515COL</td>
<td>41,110,650</td>
<td>17,017,608</td>
<td>94.71</td>
<td>100bp</td>
</tr>
<tr>
<td>Stigmas</td>
<td>Mature</td>
<td>NA1506STI</td>
<td>39,281,658</td>
<td>17,878,354</td>
<td>94.88</td>
<td>100bp</td>
</tr>
<tr>
<td>Pollen tubes</td>
<td>Germinated on pollen germination medium</td>
<td>NA1507POL</td>
<td>41,692,244</td>
<td>22,451,272</td>
<td>96.83</td>
<td>100bp</td>
</tr>
<tr>
<td>Styles</td>
<td>Without pollination</td>
<td>NA1508SNE</td>
<td>45,428,336</td>
<td>21,445,272</td>
<td>94.64</td>
<td>100bp</td>
</tr>
<tr>
<td></td>
<td>Pollinated with non-self pollen</td>
<td>NA1509STO</td>
<td>39,071,214</td>
<td>18,816,496</td>
<td>94.29</td>
<td>100bp</td>
</tr>
<tr>
<td></td>
<td>Pollinated with self-pollen</td>
<td>NA1510STS</td>
<td>50,505,064</td>
<td>23,107,862</td>
<td>95.20</td>
<td>100bp</td>
</tr>
<tr>
<td>Nectaries</td>
<td>Mature</td>
<td>NA1511NEC</td>
<td>55,777,620</td>
<td>23,403,620</td>
<td>92.92</td>
<td>100bp</td>
</tr>
<tr>
<td>Anthers</td>
<td>Mature</td>
<td>NA1512ANT</td>
<td>41,480,422</td>
<td>18,767,454</td>
<td>96.22</td>
<td>100bp</td>
</tr>
<tr>
<td>Ovaries</td>
<td>Mature</td>
<td>NA1513OVA</td>
<td>39,608,326</td>
<td>17,426,638</td>
<td>94.42</td>
<td>100bp</td>
</tr>
<tr>
<td>Pedicels</td>
<td>No treatment</td>
<td>NA1514PED</td>
<td>41,520,690</td>
<td>19,508,002</td>
<td>95.14</td>
<td>100bp</td>
</tr>
<tr>
<td>Flowers</td>
<td>Fully opened</td>
<td>NA1516OFL</td>
<td>43,791,980</td>
<td>18,550,688</td>
<td>94.87</td>
<td>100bp</td>
</tr>
<tr>
<td></td>
<td>Buds</td>
<td>NA1517FLB</td>
<td>45,864,746</td>
<td>22,167,496</td>
<td>95.47</td>
<td>100bp</td>
</tr>
<tr>
<td>Leaves</td>
<td>Leave samples that were collect at different time points</td>
<td>NA1821CTN</td>
<td>37,692,054</td>
<td>36,884,606</td>
<td>95.44</td>
<td>100bp</td>
</tr>
</tbody>
</table>
Table S6. Annotation of nicotine biosynthesis genes and number of motifs in the 2kb upstream region. TPM: transcript per million.

<table>
<thead>
<tr>
<th>Gene ID in N. attenuata</th>
<th>Gene name</th>
<th># of G-box motifs</th>
<th># of GCC motifs</th>
<th>Root expression (TPM)</th>
<th>Reference sequences in database</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIAT7_g12091</td>
<td>A02</td>
<td>5</td>
<td>2</td>
<td>3640.36</td>
<td>DW001381</td>
<td>(61)</td>
</tr>
<tr>
<td>NIAT7_g36700</td>
<td>Q5</td>
<td>2</td>
<td>2</td>
<td>3311.29</td>
<td>AF154657</td>
<td>(61)</td>
</tr>
<tr>
<td>NIAT7_g25234</td>
<td>OPT2</td>
<td>3</td>
<td>2</td>
<td>2816.12</td>
<td>AJ748263.1</td>
<td>(62)</td>
</tr>
<tr>
<td>NIAT7_g05809</td>
<td>ODC2</td>
<td>2</td>
<td>0</td>
<td>3866.97</td>
<td>AF127242.1</td>
<td>(61)</td>
</tr>
<tr>
<td>NIAT7_g61142</td>
<td>PMT1.1</td>
<td>4</td>
<td>3</td>
<td>4457.42</td>
<td>D285506.1</td>
<td>(63)</td>
</tr>
<tr>
<td>NIAT7_g05615</td>
<td>PMT1.2</td>
<td>4</td>
<td>4</td>
<td>1806.61</td>
<td>D285506.1</td>
<td>(63)</td>
</tr>
<tr>
<td>NIAT7_g14063</td>
<td>MPO1</td>
<td>3</td>
<td>3</td>
<td>454.52</td>
<td>AB289451.1</td>
<td>(64)</td>
</tr>
<tr>
<td>NIAT7_g17187</td>
<td>BBL2.1</td>
<td>2</td>
<td>2</td>
<td>275.14</td>
<td>AB604219.1</td>
<td>(65)</td>
</tr>
<tr>
<td>NIAT7_g00769</td>
<td>BBL2.2</td>
<td>4</td>
<td>3</td>
<td>956.46</td>
<td>AB604219.1</td>
<td>(65)</td>
</tr>
<tr>
<td>NIAT7_g09977</td>
<td>A622</td>
<td>4</td>
<td>0</td>
<td>2258.39</td>
<td>AB445396.1</td>
<td>(66)</td>
</tr>
<tr>
<td>NIAT7_g00125</td>
<td>DAO</td>
<td>1</td>
<td>4</td>
<td>12.01</td>
<td>XM_009602560</td>
<td>-</td>
</tr>
<tr>
<td>NIAT7_g01601</td>
<td>SPDS1</td>
<td>1</td>
<td>0</td>
<td>220.22</td>
<td>NM_001302585.1</td>
<td>(67)</td>
</tr>
<tr>
<td>NIAT7_g11058</td>
<td>MPO2</td>
<td>2</td>
<td>0</td>
<td>12.86</td>
<td>AB289457</td>
<td>(64)</td>
</tr>
<tr>
<td>NIAT7_g29386</td>
<td>A01</td>
<td>3</td>
<td>0</td>
<td>8.21</td>
<td>XM_009759430</td>
<td>-</td>
</tr>
<tr>
<td>NIAT7_g29483</td>
<td>BBL-like</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>AB604221</td>
<td>(65)</td>
</tr>
<tr>
<td>NIAT7_g34686</td>
<td>ODC1</td>
<td>0</td>
<td>2</td>
<td>11.27</td>
<td>XM_009771959</td>
<td>-</td>
</tr>
<tr>
<td>NIAT7_g58606</td>
<td>OPT1</td>
<td>1</td>
<td>1</td>
<td>29.29</td>
<td>AJ748262.1</td>
<td>(62)</td>
</tr>
</tbody>
</table>
Table S7. Primers used for validating 2kb upstream sequence regions of selected genes. While direct PCR amplifications worked well for most of the genes using primers directly binding to beginning of the protein coding region and the end of upstream 2kb fragments, nested PCR amplifications were required for four genes, likely due to the presence of repetitive sequences present in these genes.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
<th>PCR amplification</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIA07 _p00125</td>
<td>CTCATCAAAAAGTGAGAATCCGATAGAG</td>
<td>GAGGAGAAAGAACCCGCTGCCTTCCCTAGAG</td>
<td>One step amplification</td>
</tr>
<tr>
<td>NIA07 _p01601</td>
<td>GTGGTGTTTATAGAAATATGAGAACCAAC</td>
<td>CTTTGTTCATTCTGTTGTCGGTGGCG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p04912</td>
<td>CATGGGCGGGAGACATACCAAAACAGTTGAG</td>
<td>GAAGGCGCTGCACTACAGACTGCAAGC</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p05615</td>
<td>ATCCATTCTGGCTGACCTATTCATTCTACAG</td>
<td>TAGACCCATTGCTGTTGGGGAATAGAC</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p05809</td>
<td>ATCAGACCCAGACAAATAGTTATGAGCGGC</td>
<td>GAATGCGCTGCGGCTGACCTACGAAAGAAG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p09778</td>
<td>CTCCAGTTTTTCCTGACATGAGTTTCTAGCG</td>
<td>CAAGGCTCTGCTCACTTACAGACTGCAAGC</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p12091</td>
<td>GAGTGGAATTGTCCTCATAGACTGCAAG</td>
<td>CTGTGCACATCTGACGAGGATACAGCAG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p14063</td>
<td>CCCCCCTGCTGGTCATTGACCTATTGCGG</td>
<td>GTGCCGCTCATTGCTGATTAGTGAGGCC</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p20325</td>
<td>CATGAAATAGTGAGCAAAATAAGGGCTTG</td>
<td>CTCCAGGAATAGAGAAGAGGATACAGCAG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p20333</td>
<td>CTATCTGGCTACGGAACTGGGTTTGTGAG</td>
<td>GTCTATCAAGCGCTCTGACAAAAAGGAGAAC</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p21803</td>
<td>CAATGTCCTCTGACCTCGGTATAGACAG</td>
<td>CACTTTCACCTCGTGGGCTCTTACGATC</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p25524</td>
<td>CGTGTAAGATATCTGAAATGAGAAATCTC</td>
<td>GACCCTAGATATCGGAGTGGCTGAGGAG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p32818</td>
<td>GGGAGGAGTAGCTGAGAGGAAATACG</td>
<td>GACCCTAGATATCGGAGTGGCTGAGGAG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p29386</td>
<td>ATGGCGTGCTAGAAGACCTAGATGAGC</td>
<td>CGGCCCTGAGACAAAGCTTGATGAGCTG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p29883</td>
<td>GTAAAGAAGAAGCCTACCAAGAGGTG</td>
<td>GAAATGAGGCAGAACTGGAAGAAACAGATG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p32622</td>
<td>GATATCGGCGTTTATAGTAGGGAATGAAAG</td>
<td>GCACAGTGGCACTGAGCTTTCTCGCG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p34686</td>
<td>GACTCGGAGGGAAATAGCTCTCCTAAATAC</td>
<td>GGATACATAGCTACCGTAAGCTGAGTTG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p36700</td>
<td>GAGAAGAACATGACTTACAGAAGAGACC</td>
<td>GAAGATGACATAGCTACAGAAGAGACC</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p38208</td>
<td>CAGAGATGGATTCTAATGATCCCTTGT</td>
<td>CAAGGAGATGATATATCAAGAAAAGGAGAGA</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p38606</td>
<td>CTCCACACTCCAGAAGCTGCAATAGG</td>
<td>GTAATCTGCAAGATGGATGACTTAGTACGTTG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p11442</td>
<td>GCGCGCTGCTGACGAGGAGGAAATACG</td>
<td>GCATATGATGATGATGATGATGATGATG</td>
<td></td>
</tr>
<tr>
<td>NIA07 _p14350</td>
<td>GTGAGAGTTGCTGATATGAGCTCGATGAG</td>
<td>GTGAGAGTTGCTGATATGAGCTCGATGAG</td>
<td>1\textsuperscript{st} nested PCR</td>
</tr>
<tr>
<td>NIA07 _p00769</td>
<td>CATACCATCAGACTCGGAGGTTGAGG</td>
<td>CCTTCATGATGGCATAGATAGATGATG</td>
<td>2\textsuperscript{nd} nested PCR</td>
</tr>
<tr>
<td>NIA07 _p10588</td>
<td>GAGTCAATGCTGTCTTATCCCTCTTAC</td>
<td>AAGATGCGATCTGATAGCTGATGATG</td>
<td>1\textsuperscript{st} nested PCR</td>
</tr>
<tr>
<td>NIA07 _p34938</td>
<td>TGGAATATTATTACCATGACTTACCTTGT</td>
<td>GTGAGAGTTGCTGATATGAGCTCGATGAG</td>
<td>1\textsuperscript{st} nested PCR left</td>
</tr>
<tr>
<td>NIA07 _p17187</td>
<td>CAACTCTATACATATTAGTCATAGATGAG</td>
<td>CCGTAGCGATGGCATAGATGATGATG</td>
<td>2\textsuperscript{nd} nested PCR right</td>
</tr>
<tr>
<td>NIA07 _p09777</td>
<td>CTTTTGCTGCTGATGATTCTGACCTTT</td>
<td>GTGAGAGTTGCTGATATGAGCTCGATGAG</td>
<td>1\textsuperscript{st} nested PCR</td>
</tr>
<tr>
<td>NIA07 _p17187</td>
<td>CTTTTGCTGCTGATGATTCTGACCTTT</td>
<td>GTGAGAGTTGCTGATATGAGCTCGATGAG</td>
<td>1\textsuperscript{st} nested PCR left</td>
</tr>
<tr>
<td>NIA07 _p09777</td>
<td>CTTTTGCTGCTGATGATTCTGACCTTT</td>
<td>GTGAGAGTTGCTGATATGAGCTCGATGAG</td>
<td>1\textsuperscript{st} nested PCR left</td>
</tr>
</tbody>
</table>

1146
Table S8. Prediction of miRNA targeting sites on the 2kb upstream region of nicotine biosynthesis genes and their ancestral or non-root specific expressed copies. The overlap between predicted miRNA targeting sites and TE insertions are also provided in the last column. The positions indicate the seeds of predicted miRNA targeting sites.

<table>
<thead>
<tr>
<th>GeneID</th>
<th>Start</th>
<th>End</th>
<th>miRNA</th>
<th>Seed length</th>
<th>GeneName</th>
<th>Root-specific expression</th>
<th>Overlap with TE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIATv7_g00769</td>
<td>-1937</td>
<td>-1929</td>
<td>NIAATr2_miR77253p.2</td>
<td>8</td>
<td>BBL2.2</td>
<td>YES</td>
<td>rnd-5_family-2472: DNA/hAT-Ao</td>
</tr>
<tr>
<td>NIATv7_g00769</td>
<td>-1228</td>
<td>-1219</td>
<td>NIAATr2_miR1429.5p</td>
<td>9</td>
<td>BBL2.2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g00769</td>
<td>-1233</td>
<td>-1225</td>
<td>NIAATr2_miR5164</td>
<td>8</td>
<td>BBL2.2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g05615</td>
<td>-36</td>
<td>-28</td>
<td>NIAATr2_miR7504</td>
<td>8</td>
<td>PMT1.2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g05615</td>
<td>-1499</td>
<td>-1491</td>
<td>NIAATr2_miR7711.5p</td>
<td>8</td>
<td>PMT1.2</td>
<td>YES</td>
<td>rnd-5_family-822: LTR/Gypsy</td>
</tr>
<tr>
<td>NIATv7_g05615</td>
<td>-737</td>
<td>-729</td>
<td>NIAATr2_miR5163.3p</td>
<td>8</td>
<td>PMT1.2</td>
<td>YES</td>
<td>rnd-5_family-58: LTR</td>
</tr>
<tr>
<td>NIATv7_g05809</td>
<td>-1865</td>
<td>-1857</td>
<td>NIAATr2_miR1429.5p</td>
<td>8</td>
<td>ODC2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g05809</td>
<td>-1450</td>
<td>-1442</td>
<td>NIAATr2_miR162</td>
<td>8</td>
<td>ODC2</td>
<td>YES</td>
<td>rnd-4_family-757: LINE/RTE-BovB</td>
</tr>
<tr>
<td>NIATv7_g05809</td>
<td>-1865</td>
<td>-1857</td>
<td>NIAATr2_miR161</td>
<td>8</td>
<td>ODC2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g05809</td>
<td>-106</td>
<td>-95</td>
<td>NIAATr2_miR9487</td>
<td>11</td>
<td>ODC2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g09777</td>
<td>-977</td>
<td>-968</td>
<td>NIAATr2_miR1066</td>
<td>9</td>
<td>A622</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g09777</td>
<td>-688</td>
<td>-680</td>
<td>NIAATr2_miR9616</td>
<td>8</td>
<td>A622</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g12091</td>
<td>-724</td>
<td>-716</td>
<td>NIAATr2_miR8015.3p</td>
<td>8</td>
<td>AO2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g12091</td>
<td>-1092</td>
<td>-1083</td>
<td>NIAATr2_miR398.3p</td>
<td>9</td>
<td>AO2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g14063</td>
<td>-1219</td>
<td>-1211</td>
<td>NIAATr2_miR6444</td>
<td>8</td>
<td>MPO2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g14063</td>
<td>-780</td>
<td>-772</td>
<td>NIAATr2_miR8015.3p</td>
<td>8</td>
<td>MPO2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g17187</td>
<td>-1700</td>
<td>-1692</td>
<td>NIAATr2_miR5497</td>
<td>8</td>
<td>BBL2.1</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g23254</td>
<td>-793</td>
<td>-785</td>
<td>NIAATr2_miR6161</td>
<td>8</td>
<td>QPT2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g23254</td>
<td>-1559</td>
<td>-1551</td>
<td>NIAATr2_miR156.5p</td>
<td>8</td>
<td>QPT2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g23254</td>
<td>-940</td>
<td>-932</td>
<td>NIAATr2_miR5750</td>
<td>8</td>
<td>QPT2</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g36700</td>
<td>-1075</td>
<td>-1067</td>
<td>NIAATr2_miR1614</td>
<td>8</td>
<td>QS</td>
<td>YES</td>
<td>rnd-4_family-1369: DNA</td>
</tr>
<tr>
<td>NIATv7_g41142</td>
<td>-36</td>
<td>-28</td>
<td>NIAATr2_miR7504</td>
<td>8</td>
<td>PMT1.1</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g41142</td>
<td>-1042</td>
<td>-1032</td>
<td>NIAATr2_miR1429.5p</td>
<td>10</td>
<td>PMT1.1</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g60125</td>
<td>-1151</td>
<td>-1143</td>
<td>NIAATr2_miR7504</td>
<td>8</td>
<td>DAO</td>
<td>NO</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g61601</td>
<td>-1138</td>
<td>-1130</td>
<td>NIAATr2_miR5750</td>
<td>8</td>
<td>SPOSI</td>
<td>NO</td>
<td></td>
</tr>
<tr>
<td>NIATv7_g64912</td>
<td>-1586</td>
<td>-1578</td>
<td>NIAATr2_miR6444</td>
<td>8</td>
<td>ADC2</td>
<td>NO</td>
<td>rnd-1_family-18: LTR/Copita</td>
</tr>
<tr>
<td>NIATv7_g64912</td>
<td>-1190</td>
<td>-1182</td>
<td>NIAATr2_miR8015.3p</td>
<td>8</td>
<td>ADC2</td>
<td>NO</td>
<td>rnd-4_family-175: DNA</td>
</tr>
<tr>
<td>Gene</td>
<td>Start (bp)</td>
<td>End (bp)</td>
<td>Genomic Feature</td>
<td>Gene Name</td>
<td>Expression</td>
<td>Function</td>
<td>Notes</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>----------</td>
<td>---------------------</td>
<td>-----------</td>
<td>------------</td>
<td>--------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>NIATv7_g105934</td>
<td>-67</td>
<td>-59</td>
<td>NIATTr2_mIR5750</td>
<td>ADC1</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIATv7_g11058</td>
<td>-215</td>
<td>-207</td>
<td>NIATTr2_mIR8007.5p</td>
<td>MPO1</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIATv7_g11058</td>
<td>-482</td>
<td>-474</td>
<td>NIATTr2_mIR172.3p</td>
<td>MPO1</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIATv7_g29386</td>
<td>-1738</td>
<td>-1730</td>
<td>NIATTr2_mIR86667</td>
<td>AO1</td>
<td>NO</td>
<td></td>
<td>Rnd-4_family-391: DNA</td>
</tr>
<tr>
<td>NIATv7_g29483</td>
<td>-1830</td>
<td>-1821</td>
<td>NIATTr2_mIR1429.5p</td>
<td>BBL-like</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIATv7_g29483</td>
<td>-825</td>
<td>-817</td>
<td>NIATTr2_mIR408.5p</td>
<td>BBL-like</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIATv7_g29483</td>
<td>-1830</td>
<td>-1821</td>
<td>NIATTr2_mIR6161</td>
<td>BBL-like</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIATv7_g34398</td>
<td>-982</td>
<td>-974</td>
<td>NIATTr2_mIR7725.3p2</td>
<td>SPDS2</td>
<td>NO</td>
<td></td>
<td>Rnd-4_family-1284: DNA/TeMar-Stowaway</td>
</tr>
<tr>
<td>NIATv7_g34398</td>
<td>-695</td>
<td>-686</td>
<td>NIATTr2_mIR6161</td>
<td>SPDS2</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIATv7_g34686</td>
<td>-1228</td>
<td>-1220</td>
<td>NIATTr2_mIR3069</td>
<td>ODC1</td>
<td>NO</td>
<td></td>
<td>Rnd-4_family-173: DTT-NIC1</td>
</tr>
<tr>
<td>NIATv7_g34686</td>
<td>-447</td>
<td>-439</td>
<td>NIATTr2_mIR6164</td>
<td>ODC1</td>
<td>NO</td>
<td></td>
<td>Rnd-2_family-51: LTR/Gypsy</td>
</tr>
<tr>
<td>NIATv7_g58606</td>
<td>-1401</td>
<td>-1392</td>
<td>NIATTr2_mIR426</td>
<td>OPT1</td>
<td>NO</td>
<td></td>
<td>Rnd-5_family-192: LTR/Gypsy</td>
</tr>
<tr>
<td>NIATv7_g58606</td>
<td>-1962</td>
<td>-1954</td>
<td>NIATTr2_mIR4496</td>
<td>OPT1</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6. Supplemental Figures

Fig. S1. Distribution of Bio-Nano molecule length. Left panel indicates the size distribution of the BioNano molecule length, right panel shows an example of the BioNano output image.
Fig. S2. Anchoring of scaffolds to the linkage map. In total, 12 linkage groups were constructed. Linkage map and anchored pseudo-chromosomes are shown on left and right sides, respectively.
Fig. S3. Workflow used to assemble the genome of *N. attenuata*. Data used for *N. attenuata* assembly are shown in the rounded boxes. Software and assembly processes are shown on the left and right side of each arrow, respectively.
Fig. S4. Gene family size evolution among 11 plant genomes. The number of all duplication events (above branch nodes) from 11 different plant species. The duplication events were estimated from phylogenetic trees constructed from homolog groups. Only branches that have approximated Bayes branch support values greater than 0.9 are presented. Known whole genome triplication (red stars) and duplication (green stars) events are shown.
**Fig. S5. Evolution of threonine deaminase (TD).** **A)** molecular function of TD, which converts threonine to α-ketobutyrate, the substrate for the biosynthesis of isoleucine. **B)** syntenic information of TD in grape and tomato genomes. Upper panel shows the syntenic region between grape and tomato chromosome 9. Two copies of TD were found on the tomato chromosome 10 (between 61.8Mb- 63.8Mb), and one copy was found on the tomato chromosome 9, which was reverse duplicated from chromosome.
10. **C and D**. Phylogenetic tree of the TD family (C) and a simplified model (D) show the evolutionary history of TD. Numbers on each branch indicate the approximate Bayes branch supports. Local duplication and whole genome duplication events are shown as purple and green circles, respectively. **E** and **F** expression of four TDs among different tissues of *N. attenuata* (E), and among different time points after *M. sexta* OS elicitation (F) in leaves. In (F), the expression is shown as mean expression and standard error (from three biological replicates).
Fig. S6. Distribution of 4dTV is consistent with the hypothesis that *Nicotiana* and *Solanum* share a genome-wide duplication event. The distributions of fourth-fold degenerate sites (4dTV) between duplicated paralogs within and orthologs between genomes are shown. The comparison *N. attenuata* with tomato reflects the speciation events between *Nicotiana* and *Solanum*. Within genome comparisons show the divergence between duplicated gene pairs and thus reflects genome-wide duplication events.
Fig. S7. Circos plot of the 12 assembled pseudo-chromosomes. Ribbons indicate the 22 syntenic blocks, each of which contains at least 20 genes. A) Ka/Ks of genes between N. attenuata and N. obtusifolia. Red dots indicate the value greater than 1. B) M. sexta oral secretion induced gene expression changes in N. attenuata leaves. Each bar indicates log2 fold change of each gene. C) heatmap shows the distribution of transposable elements within a 500kb sliding window. Red and blue indicate high and low, respectively. D) number of genes in 500kb sliding windows.
Fig. S8. Evolution of ornithine decarboxylases (ODC). A) Phylogenetic tree of ODC among different plants. An ancient duplication event occurred before the divergence between Vitaceae and Solanaceae. The number on the branch shows the approximate Bayes branch support. B) syntenic information between the homologs of ODC1 and ODC2 in tomato. No clear signature of synteny was found. C) a dot plot depicts the sequence similarity of CDS and the 2kb upstream region between OCD1 and ODC2 in N. attenuata. D) detailed annotation of TE and transcription factor binding motifs. Light blue regions indicate the TE families annotated from RepeatMasker.
Fig. 89. Evolution of N-putrescine methyltransferases (PMT). A) Phylogenetic tree of PMT and SPERMINIDE SYNTHASE (SPDS) in six plant species. The number on the branch shows the approximate
Bayes branch support. The closest homolog from *Arabidopsis* and grape were considered as outgroups. **B)** and C) simplified evolutionary model of SPDS and *PMT* in tomato and *Nicotiana*. The syntenic information used to construct these models is from tomato. **D)** a simplified schematic representation of *SPDS* and *PMT* functions. **E)** and **F)** detailed annotation of TE and transcription factor binding motifs of *NaPMT1.1* (E) and *NaPMT1.2* (F). Light blue regions indicate the TEs annotation from RepeatMasker, and pink rounded rectangles depict the motif sequences and their 150 bp flanking region that shows significant homology to TEs (e-value < 1e-5).
Fig. S10. Evolution of N-methylputrescine oxidases (MPO). A) the phylogenetic tree of MPO and DAO in 11 plant species. MPO evolved from the duplication of DAO, likely before the divergence between Phrymaceae (Lamiales) and Solanaceae. The duplication between MPO1 and DAO2 are shared among different Solanaceae species. B) the homologues of MPO1 and DAO from tomato are in a syntenic block. C) two dot plots depict the sequence similarity of protein coding sequences and 2kb upstream regions between MPO1 and DAO (left) and between MPO1 and DAO2 from N. attenuata (right). D) detailed annotation of TE and transcription factor binding motifs. Pink rounded rectangles depict the motif sequences and their 150 bp flanking region that show significant homology to TEs (e-value < 1e-5).
**Fig. S11. Evolution of aspartate oxidases (AO).** A) Phylogenetic tree of AO among 11 plants. Both *N. attenuata* and *N. obtusifolia* have two copies of AO. The number on the branch shows the approximate Bayes branch support. B) and C) the dot plot of CDS sequence together with the 2kb upstream region between *N. attenuata* AO1 and AO2 (B), and between AO2 from *N. attenuata* and *N. obtusifolia* (C). D) detailed annotation of TE and transcription factor binding motifs. Light blue region indicates the TEs annotated from RepeatMasker, and pink rounded rectangles depict the motifs sequences and their 150 bp flanking region that show significant homology to TEs (e-value < 1e-5).
**Fig. S12. Evolution of quinolinic acid phosphoribosyltransferases (QPT).** A) The phylogenetic tree of QPT among 13 plants. All *Nicotiana* species have two copies of QPT, except *N. obtusifolia*, which is likely due to incomplete genome assembly. The number on the branch shows the bootstrap value. B and C) Dot plots depicting the sequence similarity of CDS sequence together with the 2kb upstream region of QPT1 and QPT2 in *N. attenuata* (B) and QPT2 between *N. attenuata* and *N. obtusifolia*. D) Detailed annotation of TE and transcription factor binding motifs. Light blue regions indicate the TE annotated from RepeatMasker.
Fig. S13. Evolution of berberine bridge enzyme-like proteins (BBL). A) Phylogenetic tree of BBLs among different plants. Gene structures of all BBL homologs are shown. The number on the branch shows the bootstrap value. B-E) dot plots of CDS sequence together with 2kbp upstream region between
NaBBL2.2 and NaBBL2 (B), NaBBL2.1 and NaBBL2.2 (C), NaBBL2.1 and NaBBL-like (D), NaBBL2.2 and NaBBL-like (E). F and G) detailed annotation of TE and transcription factor binding motifs of NaBBL2.1 (F) and NaBBL2.2 (G). Light blue regions indicate the TEs annotated from RepeatMasker, and pink rounded rectangles depict the motifs sequences and their 150 bp flanking region that show significant homology to TEs (e-value < 1e-5). Both NaBBL2.1 and NaBBL2.2 have an insertion of the DTT-NIC1 MITE insertion.
**Fig. S14. Evolution of MYC2 in Nicotiana.** Phylogenetic tree of MYC2 among different plants. The numbers on the branches show the bootstrap values.
Fig. S15. Distribution of expressed genes and transcripts among different tissue-specific RNA-seq libraries. A) number of expressed genes; B) number of expressed total transcripts; C) number of expressed transcripts that contain transposon sequences. The abbreviations of sample information are:

- ROT: root from plant induced by *M. sexta* OS;
- LET: leaf from plant induced by *M. sexta* OS;
- LEC: leaf from non-treated plant;
- SED: dry seeds;
- SEW: seeds treated with water;
- SES: seeds treated with liquid smoke;
- STT: stem from *M. sexta* OS-induced plant;
- COE: corollas at early developmental stage;
- COL:
corollas at late developmental stage; STI: styles; POL: pollen tubes grown in pollen germination media;
SNP: stigmas not pollinated; STO: stigmas outerossed; STS: stigmas self-pollinated; NEC: nectaries;
ANT: anthers; OVA: ovaries; PED: pedicels; OFL: open flower; FLB: flower bud.
Fig. S16. Distribution of tissue-specific expressed genes and transcripts. A) number of tissue specifically expressed genes; B) number of tissue specifically expressed transcripts.
Fig. S17. Distribution of different alternative splicing events in *N. attenuata*. AltA refers to alternative acceptor site; AltD refers to alternative donor site; IR refers to intron retention and ES refers to exon skipping.
Fig. S18. Expression of transposable elements among nine tissues. A) heatmap depicting the relative expression of 577 TE families that showed expression in at least one tissue (FPKM>1). Classification of the different TEs is shown with different colors on the left side. B) boxplots showing the expression of TEs among different tissues. Different letters indicate statistical differences (multiple comparisons after Kruskal-Wallis test $P < 0.01$) among different tissues. Numbers on each boxplot indicate median values for each tissue.
Fig. S19. Percentage of TEs induced by *M. sexta* oral secretions in *N. attenuata* leaves within a 21 h time course. Each color indicates different classes of TE families.
Fig. S20. The most abundant 20 Pfam domains in *N. attenuata.*
Fig. S21. Top 20 superfamily domains in *N. attenuata.*
Fig. S22. Distribution of gene families among four plant species.
Fig. S23. Expansion of SLF gene family in Solanaceae. Phylogenetic diagram showing the rapid gene duplication (colors) and loss (gray) of SLF subfamily members in the Solanaceae. The sequence from *Pyrus pyrifolia* was used as an outgroup. The number at each branch indicates the approximate Bayes branch supports. Red circles indicate the detected duplication shared among Solanaceae species. The color of each node indicates the species while the gene loss in a subclade is indicated by the gray regions. Nodes with only color but no gene ID indicate the gene loss detected in the given species/clade.
**Fig. S24. Expansion of the Zeatin O-glucosyltransferase-like gene family in Solanaceae.** A phylogenetic tree showing the expansion of the family. Numbers at the branches refer to the approximate Bayes branch supports. Red circles indicate the detected duplication events shared among Solanaceae species.
**Fig. S25. Expansion of a disease resistant gene family (RPM1-like) in *Nicotiana*.** A) phylogenetic tree of the gene family. This gene family was found specific to Solanaceae plants. While no duplication was found in other Solanaceae species, three duplication events were identified in *Nicotiana*. Red circles indicate gene duplication events. The numbers below the branches refer to the approximate Bayes branch supports. B) genomic localization of the four genes in *N. attenuata.*
**Fig. S26. Expansion of the purine uptake permease gene family in Nicotiana.**

**A)** phylogenetic tree of the gene family. The color indicates the species. Red circles indicate Nicotiana-specific gene duplication events. The numbers at the branches refer to the approximate Bayes branch supports. **B)** heatmap showing the expression of genes from six representative N. attenuata tissues. The orthologous genes of N. tabaccum NICOTINE UPTAKE PERMEASE1 (NUP1) is highlighted in red color. ROT: root; LET: leaf treated with M. sexta oral secretion. LEC, control untreated leaf; STT: stem from plant treated with M. sexta oral secretion in leaf; NEC: nectary; OFL: flower. Heatmap color gradient indicates the scaled log_{10} TPM value.
Fig. S27. Evolution of root-specific expression of nicotine biosynthesis genes. Heatmaps depict the expression of nicotine biosynthesis genes or their orthologs in N. attenuata, tomato and potato. The results show that PMT1, A622-like and BBL-like in tomato showed root-specific expression, although their absolute expression levels are low. It is worth noticing that the enzymatic functions of A622-like and BBL-like genes in tomato and potato remain unknown. The expression data of N. attenuata, tomato and potato were retrieved from N. attenuata datahub (http://nadh.ice.mpg.de/NaDH/), tomato eFP browser (http://bar.utoronto.ca/efp_tomato/cgi-bin/efpWeb.cgi) and potato eFP browser (http://bar.utoronto.ca/efp_potato/cgi-bin/efpWeb.cgi), respectively.
**Fig. S28. Evolution of quinolinic acid synthases (QS).** A) Phylogenetic tree of QS among 11 plants. Both *N. attenuata* and *N. obtusifolia* have one copy of QS. The number on the branch shows the approximate Bayes branch support. B) a dot plot depicts the sequence similarity of CDS sequence together with the 2kb upstream region of QS between *N. attenuata* and *N. obtusifolia*. C) detailed annotation of TE and transcription factor binding motifs. Light blue regions indicate the TEs annotated from RepeatMasker.
**Fig. S29. Evolution of A622 protein.** A) a subclade tree of the A622 and A622-like superfamily is shown among six plants. Both *N. attenuata* and *N. obtusifolia* have single copies of A622. The numbers on the branch show the approximate Bayesian support. B) the dot plot depicts the sequence similarity of protein coding sequences and 2kb upstream regions between A622 from *N. attenuata* and *N. obtusifolia*. C) detailed annotation of TE and transcription factor binding motifs. Light blue regions indicate the TEs annotated from RepeatMasker, and pink rounded rectangles depict the motifs sequences and their 150 bp flanking region that show significant homology to TEs (e-value < 1e-5).
7. Captions for supplementary datasets S1 to S9

**Supplementary Dataset 1. Summary information on MITE families in Nicotiana and their homologous families in Solanum genomes.** Different superfamilies are represented by different codes. DTA for Tcl/Mariner, DTA for hAT, DTH for PIF/Harbinger. For tomato and potato, the original MITE IDs are shown in the last column. The biggest MITE family in Nicotiana, DTT-NIC1, is highlighted in bold font.

**Supplementary Dataset 2. Annotation of smRNAs in N. attenuata.**

**Supplementary Dataset 3. Expression of genes among 21 RNA-seq libraries.** The expression numbers are TPM values. The abbreviation of sample information are: ROT: root from plant induced by *M. sexta* OS; LET: leaf from plant induced by *M. sexta* OS; LEC: leaf from non-treated plant; SED: dry seeds; SEW: seeds treated with water; SES: seeds treated with liquid smoke; STT: stem from *M. sexta* OS-induced plant; COE: corollas at early developmental stage; COL: corollas at late developmental stage; STI: styles; POL: pollen tubes grown in pollen germination media; SNP: stigmas not pollinated; STO: stigmas outcrossed; STS: stigmas self-pollinated; NEC: nectaries; ANT: anthers; OVA: ovaries; PED: pedicels; OFL: open flower; FLB: flower bud; CTN: leaf samples collected at different time points pooled together. Tissue specificity was calculated using the $\tau$ index among all tissues except CTN.

**Supplementary Dataset 4. Expression of assembled transcripts among 21 RNA-seq libraries.**

The expression numbers are TPM values. The abbreviation of sample information are: ROT: root from plant induced by *M. sexta* OS; LET: leaf from plant induced by *M. sexta* OS; LEC: leaf from non-treated plant; SED: dry seeds; SEW: seeds treated with water; SES: seeds treated with liquid smoke; STT: stem from *M. sexta* OS-induced plant; COE: corollas at early developmental stage; COL: corollas at late developmental stage; STI: styles; POL: pollen tubes grown in pollen germination media; SNP: stigmas not pollinated; STO: stigmas outcrossed; STS: stigmas self-pollinated; NEC: nectaries; ANT: anthers; OVA: ovaries; PED: pedicels; OFL: open flower; FLB: flower bud; CTN: leaf samples collected at different time points pooled together. Tissue specificity
and transcripts of which at least 50bp were annotated as TE are also indicated. Tissue specificity was calculated using the τ index among all tissues except CTN.

**Supplementary Dataset 5. Expression of microarray probes that are both induced by simulated herbivory by application of *M. sexta* oral secretion and mapped to transposable elements of *N. attenuata*.** SL: systemic untreated leaf; TL: treated leaf; RT: systemic untreated root.

**Supplementary Dataset 6. KEGG ortholog (KO) annotation of *N. attenuata* genes.**

**Supplementary Dataset 7. List of genes annotated as protein kinases and transcription factors.**

**Supplementary Dataset 8. List of recent duplication events of *N. attenuata* genes and their duplicated copies.** Annotations of the duplication that the gene was most recently involved, and the functional annotation of gene based on Blast2GO are also indicated. Detected duplication times are: *N. attenuata* specific; shared among *Nicotiana*; shared among Solanaceae; shared with *M. guttatus*; shared among core eudicots. Inferred duplication types are: genome-wide duplications, tandem duplications.

**Supplementary Dataset 9. Gene families that expanded significantly in *Nicotiana*.**
8. References


5 Manuscript II
“Evolution of herbivore-induced early defense signaling was shaped by gene duplications and transposons in *Nicotiana*”

(Zhou W, Brockmöller T, Ling Z, Omdahl A, Baldwin IT, Xu S)
Evolution of herbivore-induced early defense signaling was shaped by genome-wide duplications in Nicotiana

Wenwu Zhou¹, Thomas Brockmüller¹, Zihao Ling¹, Ashton Omdahl¹,², Ian T Baldwin¹, Shuqing Xu¹*

¹Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Jena, Germany; ²Brigham Young University, Provo, United States

Abstract

Herbivore-induced defenses are widespread, rapidly evolving and relevant for plant fitness. Such induced defenses are often mediated by early defense signaling (EDS) rapidly activated by the perception of herbivore associated elicitors (HAE) that includes transient accumulations of jasmonic acid (JA). Analyzing 60 HAE-induced leaf transcriptomes from closely-related Nicotiana species revealed a key gene co-expression network (M4 module) which is co-activated with the HAE-induced JA accumulations but is elicited independently of JA, as revealed in plants silenced in JA signaling. Functional annotations of the M4 module were consistent with roles in EDS and a newly identified hub gene of the M4 module (NlLRK1) mediates a negative feedback loop with JA signaling. Phylogenomic analysis revealed preferential gene retention after genome-wide duplications shaped the evolution of HAE-induced EDS in Nicotiana. These results highlight the importance of genome-wide duplications in the evolution of adaptive traits in plants.

DOI: 10.7554/eLife.19531.001

Introduction

Induced defense is widespread in plants and can improve the fitness of plants under herbivore attack (Baldwin, 1998; Kessler et al., 2004). Many plants recognize and distinguish the damage caused by feeding insects from mechanical damage by perceiving herbivore-associated elicitors (HAE) to induce rapid early defense signaling (EDS) that includes the accumulation of jasmonic acid (JA) and its derivatives, phytohormones that play a central role in the activation of induced defenses (Erb et al., 2012; Howe and Jander, 2008; Wu and Baldwin, 2010). Increases or decreases in leaf JA concentrations can directly activate or impair induced anti-herbivore defenses, respectively (Farmer and Ryan, 1992; Kessler et al., 2004; Wu and Baldwin, 2010), highlighting the importance of JA accumulation for induced defenses. However, increased JA levels can also reduce plant fitness due to the physiological and ecological costs of defense elicitation when defenses are not needed (Baldwin and Hamilton, 2000; Glawe et al., 2003; Heil and Baldwin, 2002; van Dam and Baldwin, 1998). For example, in Nicotiana attenuata, an increase in endogenous JA levels by supplying methyl-jasmonic acid (MeJA) reduced plant fitness by 26% when plants were protected from herbivore attack (Baldwin, 1998). Thus induced JA accumulations can result in net fitness gains or losses depending on the cost/benefit ratio of induced defenses, which varies among attacking herbivore species and environmental conditions. Therefore, a robust and complex signaling network that regulates and fine-tunes induced JA biosynthesis, metabolism and JA-dependent induced downstream defenses is essential for plants to realize their fitness optima.

Using reverse genetics, such as RNA-inference (RNAi) and virus induced gene silencing (VIGS), several genes that are rapidly induced by HAE were found to regulate JA biosynthesis and metabolism in plants, particularly in the wild tobacco Nicotiana attenuata which has been established as an...
eLife digest A variety of different insects feed on plants and these insects often produce molecules known as elicitors that the plants can recognize. This triggers a sophisticated suite of defenses in the plant that can either deter feeding by the insects, or help the plants endure the attack. The elicitors stimulate the rapid accumulation of a plant hormone called jasmonic acid, which in turn activates the defense responses. However, high levels of jasmonic acid can also reduce the ability of the plants to survive and reproduce by activating plant defenses when they are not needed. Therefore, plants need to regulate the signaling networks that control defense so that jasmonic acid only accumulates when the benefit of fighting the insect outweighs the cost of producing the defenses.

The costs and benefits of defense responses vary among different insects and environmental conditions, which has made it difficult to study how plants regulate defense signaling networks. To address this question, Zhou et al. investigated the activities of genes in six species of tobacco plant after they have been exposed to different insect elicitors.

The experiments identified a network of genes that is activated in response to elicitors and acts largely independent of jasmonic acid signaling. A newly identified gene in this network called NaLRRK1 and jasmonic acid suppress each other, suggesting that NaLRRK1 helps to regulate jasmonic acid levels. Further analysis shows that a process called genome duplication, in which all the genes in an organism are copied, has shaped the evolution of early defense signaling in Nicotiana. Many of the duplicated genes have adopted new roles and been retained in the plants. This highlights the importance of genome duplications in helping plants to adapt to their environment.

The next challenge following on from this work would be to identify what specific roles these genes play in the plants, and how they affect the ability of plants to survive insect attacks in their native habitats.

DOI: 10.7554/eLife.19531.002

ecological model system for plant-herbivore interactions (Wu and Baldwin, 2010). The HAE-regulated signaling network includes: protein kinases, such as the wounding induced protein kinase (NaWIPK) (Wu and Baldwin, 2010; Wu et al., 2007) and calcium-dependent protein kinases (NaCDPK4/5) (Yang et al., 2012), which positively and negatively regulate HAE-induced JA accumulations, respectively; transcription factors, such as NaWRKY3/6, which positively regulate HAE-induced JA accumulations (Skibbe et al., 2008); and ethylene (ET) biosynthesis and perception genes (NaETR1, NaACO and NaACS) (von Dahl et al., 2007), which crosstalk with JA-regulated downstream defense responses (Onkokesung et al., 2010; Voelckel et al., 2001), such as nicotine biosynthesis (Kahl et al., 2000; Shoji et al., 2000). While these studies have provided mechanistic insights into induced defenses, they also revealed the complexity of the HAE-induced EDS network.

A systematic investigation of its complete genetic architecture is essential to understanding the molecular mechanisms and evolution of HAE-induced EDS.

Gene duplications play a key role in network evolution (Pastor-Satorras et al., 2003; Telchmann and Babu, 2004). Duplicated genes can either be retained in the same network to increase network complexity and robustness or evolve to function in new networks through subfunctionalization and/or neofunctionalization processes (De Smet and Van de Peer, 2012; Duarte et al., 2006) that can be detected from changes in the spatiotemporal expression or protein interaction patterns of the duplicated genes. Although both gene expression and protein-protein interaction divergences between duplicated genes increase over time (Arabidopsis Interactome Mapping, 2011), several factors, such as the type of duplication and the functionality of the genes, affect the rate and extent of those divergences (Hanada et al., 2008; Rizzon et al., 2006). For instance, expression divergences between duplicated genes involved in stress responses tend to be greater than those of duplicated genes involved in developmental processes (Ha et al., 2007). While studies based on the analysis of gene ontologies and genome-wide duplications suggest that lineage-specific duplication (LD) followed by expression divergence are important for the evolution of stress responses in plants (Hanada et al., 2008; Rizzon et al., 2006), whole genome duplications (WGD)
events, which are prominent in the plant kingdom, provide a major source of duplicated genes and contribute significantly to the evolution of cellular networks, such as gene regulatory (Blanc and Wolfe, 2004), protein-protein interaction (Arabidopsis Interactome Mapping, 2011) and metabolic networks (Gachon et al., 2005; Hoffberger et al., 2013). Furthermore, duplicated copies from WGD events are more likely to be retained in a network than those from LD, especially for genes that are dosage-sensitive, such as transcription factors, and protein kinases (Arabidopsis Interactome Mapping, 2011; Birchler and Veitia, 2007; Casneuf et al., 2006; Edger and Pires, 2009; Free-ling, 2009). However, the relative contribution of WGD and LD to the evolution of HAE-induced EDS networks and the patterns of expression divergence between duplicated genes in these networks have not been studied.

Understanding the molecular mechanisms and evolution of HAE-induced EDS requires the identification of the genome-wide HAE-induced EDS networks. Because functionally related genes tend to be transcriptionally coordinated (Persson et al., 2005; Stuart et al., 2003), co-expression network analysis has been widely used to infer the function of genes and uncover biological pathways (Klie et al., 2014; Usadel et al., 2009; Yonekura-Sakakibara et al., 2008). Distinct from ‘classical’ gene expression analysis using genome-wide expression profiling of control and treated samples to identify ‘up’ or ‘down’ regulated genes, co-expression network analysis uses expression measurements from a large number of samples that vary in their genotype, treatment, tissue or sampling time to enhance the statistical power of the analysis (Zhang and Horvath, 2005). However, due to the high specificity among tissues and treatments and the speed of the HAE-elicited responses (within 30 min) (Gulati et al., 2014, 2013; Kim et al., 2011), the general co-expression network approach that uses gene expression data from different tissues or time course experiments are not particularly useful. One solution to overcome this specificity issue is to use natural variation, such as occurs amongst closely related species or different genotypes within species, to identify co-expressed gene networks (Ardlie et al., 2015; Delker et al., 2010). In the identification of HAE-induced EDS networks, the comparison of closely related species has at least two advantages over the use of different genotypes within a species: (1) their greater genetic and phenotypic diversity (Xu et al., 2015) which increases the power of detecting co-expressed genes; (2) their divergence times are over several millions of years which allows for the identification of evolutionarily conserved co-expression networks that are likely functionally important.

Closely-related Nicotiana species within the clade of Petunioideae show highly specific HAE-induced defenses and thus provide an ideal system for identifying HAE-induced EDS networks (Xu et al., 2015). Our previous study revealed that a single HAE, such as the fatty acid-amino acid conjugate C18:3-Glu (FAC) – the most active elicitor found in the oral secretions of the Solanaceae specialist herbivore Manduca sexta (OS₃₆₆) larvae – elicits diverse defense responses among closely related Nicotiana species when added to standardized puncture wounds. In addition, a single Nicotiana species, such as N. pauciflora, showed distinct defense responses to the FAC, OS₃₆₆ and oral secretions from the generalist herbivore Spodoptera littoralis (OS₄) (Xu et al., 2015). Here we sequenced the leaf transcriptomes of six closely-related Nicotiana species from the Petunioideae clade (N. obtusifolia, N. linearis, N. acuminata, N. pauciflora, N. miersii and N. attenuata) that had been induced by three different HAEs or simply wounded (induced by wounding plus water) to characterize the HAE-induced EDS networks in Nicotiana. We compared HAE-induced transcriptomic responses among the six species and identified a co-expression gene network that represents the HAE-induced EDS in Nicotiana based on three independent lines of evidence: (1) the induction of the network correlates with variation in JA accumulations both among species treated with the same HAE and within species treated with different HAEs; (2) the induction of genes in this network that are largely not dependent on induced JA accumulations; (3) the consequences of silencing a hub gene in this network for HAE-induced JA metabolism and defenses. Analysis of the evolutionary history of all genes in the EDS network revealed that preferential gene retention after the Solanaceae whole genome triplication (WGT) event shaped the evolution of HAE-induced EDS in Nicotiana.
Results

FAC-induced early leaf transcriptomic responses are highly variable among closely related Nicotiana species

Closely related Nicotiana species showed highly divergent early transcriptomic responses within 30 min of FAC elicitation (Figure 1), consistent with observations from metabolomic and insect performance studies (Xu et al., 2015). Two species, N. obtusifolia and N. miersii, which did not respond to FAC-treatments by amplifying their wound-induced accumulations of jasmonic acid (JA) within 2 hr, showed overall little induced transcriptomic responses (Figure 1 and Figure 1—figure supplement).

Figure 1. FAC elicits divergent transcriptome responses among closely related Nicotiana species. (a) FAC-induced JA responses among six Nicotiana species. Phylogenetic tree was constructed based on orthologous genes and numbers on each branch indicates bootstrap values. X-axis indicates time after elicitation and Y-axis denotes JA concentrations. FM= fresh mass. Gray and black colored lines refer to control (wounding and water) and FAC-induced samples, respectively. Different letters indicate significance between two treatments (Student's t-test, p<0.05). (b) transcriptomic similarity between control and FAC-induced samples (30 min after elicitation) in the six species (order is same as panel a). The color gradients indicate the Pearson correlation coefficients among samples. (c) number of differentially up- and down-regulated genes after FAC elicitation in the six species (order is same as panel a). Y-axis depicts the number of genes. Each colored bar indicates a different species. Blue: N. obtusifolia, light green: N. linearis, dark green: N. attenuata, light blue: N. miersii, orange: N. acuminata, pink: N. pauciflora. d and e, Venn diagrams of up- (d) and down- (e) regulated genes in each of the six Nicotiana species. Circle size indicates the relative number of up/down regulated genes in each species. Each filled circle indicates a different species, with color code as in panel a.

DOI: 10.7554/eLife.19531.003

The following figure supplements are available for figure 1:

Figure supplement 1. Z-score of FAC-induced gene expression changes in six species.

DOI: 10.7554/eLife.19531.004

Figure supplement 2. Validations of 12 selected FAC-induced genes in N. attenuata.

DOI: 10.7554/eLife.19531.005
A co-expressed gene module is induced by FAC-elicitation but not by JA

The highly divergent FAC-induced transcriptomic responses provide an excellent opportunity to identify co-expression networks. We identified FAC-induced gene co-expression networks using the weighted gene co-expression network analysis (WGCNA) method (see details in Materials and Methods). In total, five gene modules (M1-M5) were identified using control (wounding + water) and FAC-induced gene expression profiles from all six species (Supplementary figure 2A). Among these five modules, module M4 showed the highest correlation with HAE-induced JA accumulations, a marker of induced defense signal (Figure 2B). In all four species that showed FAC-induced JA accumulations (Figure 2C), the majority of M4 module genes were also significantly induced by FAC (p<0.05 and fold change greater than 1.5, exact negative binomial test). In contrast, in the two species, N. obtusifolia and N. miersii, which did not show FAC-induced JA accumulations (Figure 2C), less than 22% of the M4 module genes were induced. The intra-modular connectivity of the M4 module, a parameter that indicates the degree of co-expression among genes in a network, was significantly higher in FAC-induced samples than in control samples (p=0.0002, Kruskal-Wallis rank sum test), consistent with the observation that FAC elicits co-expression among genes in the M4 module (Figure 2D). Furthermore, the expression kinetic analysis of the M4 module genes using a previously published microarray dataset (Kim et al., 2011) revealed that most of the M4 module genes were largely transiently expressed (Figure 2—Supplementary figure 1) after HAE-elicitation. These data suggest that the identified M4 module is likely associated with FAC-induced EDS.

More than 53% of the M4 module genes were induced in N. pauciflora at 30 min after FAC-elicitation (Figure 2B), the time point when JA was not yet induced in this species, indicating that the induction of gene module M4 is independent of or precedes that of JA. To further test this hypothesis, using the N. attenuata genome-wide microarray, we measured FAC-induced gene expression changes in JA deficient N. attenuata plants (rsAOC), in which a key JA biosynthesis gene was silenced and the induced JA levels were reduced to basal levels (Kallenbach et al., 2012). For the comparison, we also performed genome-wide microarray analysis for the same N. attenuata wild type (WT) RNA samples that were used for the RNA-seq analysis. In the WT samples, fewer FAC-induced genes were detected by the microarray (771) than by the RNA-seq analysis (1752); however, more than 81.2% of the FAC up-regulated genes identified using the microarray were also found from the RNA-seq analysis, indicating an overall consistency between RNA-seq and microarray data, and the expected higher power and sensitivity of RNA-seq in detecting differentially expressed genes. More than 87% of the M4 module genes that were induced by FAC in WT plants, both from the RNA-seq and microarray experiments, were also up-regulated in the JA-deficient rsAOC plants (Figure 3). Likewise, based on the microarray data of samples that were collected at 30 min after FAC-elicitation, the majority (85.1%) of up-regulated genes in WT plants were also up-regulated (FDR adjusted p<0.05, fold change >1.5) in the JA-deficient plants, suggesting that the FAC-induced early expression changes are largely not dependent on FAC-induced JA accumulations. Together, these data suggest that the genes of the M4 module are largely induced by FAC but not by JA.
The induction of M4 genes is associated with the specificity of HAE-induced early defense responses within species

A previous study revealed that different HAE can induce distinct defense responses within the same species (Xu et al., 2015). To understand the underlying molecular mechanisms, we additionally sequenced the transcriptomes of leaves that were induced by the oral secretions (OS) of the Solanaceae specialist herbivore M. sexta (OS_{AS}) and the generalist herbivore S. littoralis (OS_{G}) in four different Nicotiana species that showed specific responses to different HAE. This analysis revealed that the level of M4 module gene inductions correlate with the specificity of HAE-induced defense responses within a species. In N. attenuata, FAC, OS_{AS}, and OS_{G} induced similar levels of induced defense responses, and consistently, the majority of the M4 module genes were induced by all three elicitors (Figure 4 a–d). In N. pauciflora, while both FAC and OS_{AS} up-regulated a large fraction of the M4 module genes (Figure 4b) and downstream induced defense responses, OS_{G} only up-regulated 14.8% of the M4 module genes and failed to activate the downstream defense responses (Xu et al., 2015). Furthermore, while in N. obtusifolia and N. miersii, both FAC and OS_{AS} only up-regulated less than 13.9% of the M4 module genes (Figure 4b) and did not activate the downstream defenses (Xu et al., 2015), OS_{G} up-regulated 53.3% of the M4 module genes and induced
downstream defense responses in these two species. These data suggest that the induction of the M4 module genes correlates with the variation of different HAE induced defense responses within species.

Although at a global level, the FAC and OSMa induced similar levels of phytohormones and transcriptomic defense responses (Figure 4a–c), the resulting downstream defenses, such as effects on caterpillar growth rates can be different. For example, our previous study showed that larvae grew faster on leaves induced by OSMa than by FAC in both N. pauciflora and N. attenuata (Xu et al., 2015). Because FAC is a subset of the elicitors in OSMa, we reasoned that OSMa might contain other elicitors that suppress the downstream responses of JA accumulations (Xu et al., 2015). Consistent with this hypothesis, at a transcriptomic level, we found that OSMa induced a smaller number of M4 genes than did FAC in both N. attenuata and N. pauciflora (Figure 4b). In N. miersii, in which both FAC and OSMa did not induce defense responses. We further identified eight genes (Supplementary file 1B) from the M4 module that showed lower expression in response to OSMa than to FAC in both N. attenuata and N. pauciflora. Among these eight genes, one gene, NaJAR1.1 (NIAV7. g23173, Figure 4e), is a member of the jasmonic acid-amido synthetase (JAR1) gene family (Figure 4f). JAR1 catalyzes the formation of jasmoyl-isoleucine (JA-ile), a conjugate of JA that activates downstream defense responses (Kang et al., 2006; Staswick et al., 2002). In N. attenuata genome, there are three JAR1 copies that resulted from duplication events, and two of these (NaJAR4 and NaJAR6) are induced by both FAC and OSMa and are involved in the conjugation of JA to amino acids and anti-herbivore defense responses (Kang et al., 2006; Wang et al., 2007). Although the exact functions of NaJAR1.1 remain unknown, it shares more than 85% of protein sequence identity to NaJAR4 and NaJAR6 and has the conserved amino acid conjugation domain shared by all JAR1 family members, suggesting that NaJAR1.1 is also likely involved in the metabolism of JA. We hypothesize that an unknown component in OSMa, which might be used by the specialist herbivore M. sexta to suppress the expression of NaJAR1.1 in order to regulate JA metabolism and thus suppress downstream defense responses in Nicotiana.

In summary, the induction of genes in the co-expression module M4 is associated with the specificity of HAE-induced early defense responses within species, and is consistent with the notion that induction of the M4 module is important for HAE-induced defense responses in the genus Nicotiana.
Figure 4. The induction of module M4 is associated with the specificity of different HAE-induced defense responses within species. a and b, the relative JA induction (a) and proportion of genes in the M4 module (b) induced by different HAEs in four Nicotiana species. The JA induction was scaled between 0 and 1, to indicate the lowest and highest JA level induced by three different HAEs and control (WW). Each colored bar represents elicitation from different HAEs. Dark green: FAC, purple: M. sexta oral secretion (OSm), light blue: S. littoralis oral secretion (OSl). c) Venn diagrams showing the overlap among upregulated genes induced by three different HAEs within each species. Each color represents one HAE with the same color code as in panels a and b. The sizes of the circles represent the total number of genes in each group. d-g) heatmaps showing the expression of M4 gene module members in four species as induced by the three HAEs and control. The color gradient represents the relative expression value. h) OSm induced lower expression level of JAR1.1 than did FAC in three Nicotiana species. Each bar presents the average expression (TMM normalized FPKM) of JAR1.1 in each species. Each color indicates different treatments. Gray: control (wounding and water), dark green: FAC, purple: OSm. i) the phylogenetic tree showing the relationship among three paralogs of JAR1 in N. attenuata and orthologues of JAR1 in Arabidopsis thaliana (At), Vitis vinifera (V), and Solanum lycopersicum (Sl).

The M4 module represents the conserved herbivore-induced EDS network among different Nicotiana species

The specific responses induced by different HAEs within a species also provided an opportunity to further examine the conservation of the M4 module in Nicotiana. We analyzed the preservation of the M4 module in N. attenuata, N. mierii, N. pauciflora and N. obtusifolia, of which we sequenced the transcriptomes of leaves induced by different HAEs to characterize transcriptional responses. The results revealed that the 2-summary scores of the M4 module, which indicate the level of module preservation, are all above 20 (values above 10 indicates that the module is highly conserved) for all pair-wise species comparisons, suggesting the M4 module has been retained among different species (Table 1). The statistical significance of the module preservation is further supported by permutation tests (in all comparisons, p<2.2E-16). We further analyzed the sequence divergence of M4 module genes by calculating the \( \omega \) (Ka/Ks ratio) of M4 module genes shared between N. attenuata and N. obtusifolia, the two most divergent species in the dataset. The results revealed that the \( \omega \) value of most M4 module genes (94.5%) are significantly less than 1 (p<0.05, Fisher’s exact test, median \( \omega \)=0.19), indicating they were under strong purifying selection. The distribution of \( \omega \) from M4 module genes was not different from all leaf expressed genes (median \( \omega \)=0.20, p=0.18, Wilcoxon–Mann–Whitney test), suggesting that the M4 module genes were not subject to strong divergent selection between N. attenuata and N. obtusifolia. Together, these results are consistent with the hypothesis that the identified M4 module is conserved among the different Nicotiana species.
**Table 1.** M4 module is highly preserved among four studied Nicotiana species. The number in each cell refers to the z-summary score calculated using `modulePreservation` function from WGCNA package. Species in row and column indicate the reference and testing datasets, respectively. The score above 10 indicates the co-expression module is preserved, whereas the score bellow 2 indicate the module is not preserved. For all comparisons, p-values based on permutation tests are smaller than 2.2E-16.

<table>
<thead>
<tr>
<th></th>
<th><em>N. obtusifolia</em></th>
<th><em>N. attenuata</em></th>
<th><em>N. miersii</em></th>
<th><em>N. pauciflora</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>N. obtusifolia</em></td>
<td>-</td>
<td>20.5</td>
<td>18.2</td>
<td>21.0</td>
</tr>
<tr>
<td><em>N. attenuata</em></td>
<td>29.4</td>
<td>-</td>
<td>34.9</td>
<td>31.2</td>
</tr>
<tr>
<td><em>N. miersii</em></td>
<td>22.8</td>
<td>38.8</td>
<td>-</td>
<td>27.8</td>
</tr>
<tr>
<td><em>N. pauciflora</em></td>
<td>23.4</td>
<td>26.2</td>
<td>27.4</td>
<td>-</td>
</tr>
</tbody>
</table>

DOI: 10.7554/eLife.19531.010

The M4 co-expression module contains 1274 genes, which were enriched for gene ontology terms with ‘regulation of defense responses’, ‘jasmonic acid metabolism’, ‘response to insects’, and ‘protein modification’ among others (Figure 5—figure supplement 1). A majority of the JA biosynthetic genes were found in this module and their expressions were positively correlated with each other (Figure 5), indicating that the identified co-expressed genes reflect their functional relationships. Among all M4 module genes, 782 were significantly induced by FAC in the three species (*N. attenuata, N. acuminata* and *N. linearis*) which all showed JA accumulations 30 min after FAC elicitation. We infer that these 782 genes represent the core HAE-induced EDS network, which includes 75 protein kinases and 96 transcription factors (Figure 5). Previous research has shown that silencing genes in this conserved signaling network can directly affect herbivore-induced JA biosynthesis, metabolism and downstream defenses in *N. attenuata*. This includes the following protein kinase-encoding genes: NaWIPK (Wu et al., 2007), NaMPK4 (Hettenhausen et al., 2013), NaBAK1 (Yang et al., 2011) and NaCDPK4/5 (Wu et al., 2007; Yang et al., 2012), which are positive or negative regulators of JA biosynthesis and induced defense in *N. attenuata*; as well as the transcription factor, NaWRKY6, which is involved in differentiating mechanical wounding from herbivore attack and mediates plants’ herbivore-specific defenses (Skibbe and Gallis, 2008). Furthermore, several genes in this network have also been shown to be involved in phytohormone crosstalk and regulate JA-induced downstream defense responses, including NaACO2, NaACS3a and NaETR1 which are involved in ET biosynthesis and perception (von Dahl et al., 2007); NaLecRK (Gillardoni et al., 2011; von Dahl et al., 2007) that inhibits SA accumulation during herbivory and NaFTR1 that suppresses abscisic acid (ABA) metabolism after herbivore attack, which, in turn, activates JA accumulation and defenses against insect herbivores (Diniz et al., 2013).

A hub gene of the M4 module, NaLRRK1, forms a negative feedback loop with jasmonate signaling in the herbivore-induced EDS

Hub genes, which are defined as highly connected genes in the network, are often functionally important. Based on intra-modular connectivity, we identified 64 hub genes (top 5%) in the FAC-induced co-expression network, which include NaWIPK, a key positive regulator of JA biosynthesis and induced defense in *N. attenuata* (Meldau et al., 2009). To provide further mechanistic understanding of these hub genes in regulating induced defenses, we characterised an additional unknown hub gene encoding a putative leucine-rich repeat receptor kinase (NaLRRK1). The plasma membrane and nuclei localized NaLRRK1 (Figure 6a) has an N-terminal extracellular region, a single transmembrane domain, and a C-terminal cytoplasmic region. The expression of NaLRRK1 was co-upregulated with induced JA signaling among the six Nicotiana species (Figure 6b). Measuring NaLRRK1 transcripts in leaves treated with different HAEs and one pathogen-associated elicitor, flg22, revealed that NaLRRK1 is specifically induced by HAE (Figure 6c). We investigated whether HAE-induced JA signaling regulates the expression of NaLRRK1 using two different jasmonate deficient transgenic plants, in which steps in JA signaling and perception were individually silenced (Figure 6d). Consistent with the microarray results, NaLRRK1 was still significantly induced by FAC 30 min after elicitation in both of the JA-signaling deficient genotypes, revealing that JA is not
Figure 5. The co-expression network of module M4. (a) the network view of the M4 module. Each node represents a gene in the M4 module, except the filled orange node, which represents a collapsed node from a cluster. The shape of the node represents the property of the gene. Transcription factor: triangle; round rectangle: protein kinases, ellipse: other genes. The size of each node indicates their log2 fold-change after FAP induction. The color of each node represents its Mapman functional annotation. Green: signaling, yellow: transcriptional regulation, red: post translational modification, gray: programmed cell death, purple: biotic and abiotic stress responses, dark blue: transport activity, light blue: hormone metabolism, orange: others. Edges represent the connections between two genes, estimated based on their co-expression coefficient. The genes that were shown to regulate HAE-induced anti-herbivore defenses are also shown in the network. (b) the correlation among genes involved in JA biosynthesis and metabolism. The left side shows biosynthesis and metabolism of JA, right side shows the correlation among each other. Each circle indicates the pairwise correlation coefficient between two genes. The size of the circle indicates the coefficient value. Only statistically significant correlations were shown (p<0.05, Pearson’s product moment correlation test).

DOI: 10.7554/eLife.19531.011

The following figure supplement is available for figure 5:

Figure supplement 1. Gene ontology (GO) enrichment analysis of the M4 module genes.

DOI: 10.7554/eLife.19531.012
Figure 6. Jasmonate signaling suppresses the expression of NaLRRK1. (a) Subcellular localization of NaLRRK1. Nicotiana attenuata leaves were transformed with PM::CFP and NaLRRK1::YFP. After incubation for 48 hr, the transformed leaves were observed under a confocal microscope. The photographs were taken in UV light, visible light (bright field) and in combination (merged signals). Scale bar, 20 μm. (b) The transcript accumulation of LRRK1 gene in the leaves of six Nicotiana species elicited by wounding + water (W + water) and wounding + FAC (W + FAC), estimated from RNA-seq data (n=3). Asterisk indicates FDR-adjusted p value <0.05 and fold change greater than 2. c, the kinetics of NaLRRK1 transcript accumulation in N. attenuata leaves at 0, 0.5, 1 and 2 hr after treatments with different elicitors. For each treatment, 20 μL water or elicitors: FAC, oral secretions from M. sexta (OS_{ms}), S. littoralis (OS_{sl}), or flg22 was applied to the wounded leaves. Triple asterisks indicate the significant difference (p<0.01, n=3). Except flg22 treatment was with 3 replicates between treatment and control (W + water). d, a simplified model of JA biosynthesis and metabolism. The two transformed lines, in which AOC and COI were silenced respectively, are indicated: e and f, the transcript accumulation of NaLRRK1 in the two transformed lines in comparison to WT after elicitation with water (e) or FAC (f). g, the FAC induced NaLRRK1 transcript accumulation in N. attenuata leaves was suppressed by JA-Ile. N. attenuata leaves were collected at 1 hr after the induction. JA-Ile was applied in two different concentrations. For panel e, f and g, four biological replicates were used. In panel b, c, e and f, data are presented as means ± SEM. Asterisk indicates significant difference (*: p<0.05, ***: p<0.01, Student’s t-test) between treatments.

DOI: 10.7554/eLife.19531.013

The following figure supplement is available for figure 6:

Figure supplement 1. In N. attenuata, OS_{ms} induced higher NaLRRK1 transcript levels in 35S::jmt7/7-mje plants than in WT plants.

DOI: 10.7554/eLife.19531.014

required for the up-regulation of NaLRRK1. Interestingly, compared to WT plants at 1 hr, NaLRRK1 transcript levels were higher in iAOC plants – in which JA-Ile levels remain at basal levels - but were lower in iCOI plants, in which JA-Ile levels are constitutively high (Paschold et al., 2008) (Figure 6e and f). This indicates that JA-Ile levels may suppress the accumulation of NaLRRK1 transcripts. To test this hypothesis, we compared NaLRRK1 transcript accumulations in leaves in which the levels of JA-Ile were elevated by adding different amounts of JA-Ile to wounded leaves together with FAC. The results revealed that increased JA-Ile levels indeed decreased the levels of NaLRRK1 transcripts (Figure 6g). Furthermore, in the transgenic plants 35S::jmt7/7-mje, in which endogenous JA levels are redirected to MeJA resulting in lower levels of induced JA-Ile and abrogated JA-signaling compared to WT plants (Stitz et al., 2011), HAE-induced NaLRRK1 transcript accumulation was higher than in WT plants (Figure 6—figure supplement 1). These results are consistent with the hypothesis that JA-Ile negatively regulates NaLRRK1 transcript levels.

We further investigated the roles of NaLRRK1 in regulating HAE-induced defenses in N. attenuata using virus induced gene silencing (VIGS), which reduced HAE-induced NaLRRK1 transcript
abundance by more than 88% in comparison to empty vector (EV) plants (Figure 7—figure supplement 1). The levels of a precursor of JA, OPDA, were significantly increased in VIGS-NaLRRK1 plants.

Figure 7. Silencing NaLRRK1 increases FAC-induced JA biosynthesis and metabolism and downstream defenses. (a–f) the VIGS-NaLRRK1 plants have enhanced FAC-induced transcript accumulations of genes involved in JA biosynthesis and metabolism compared to EV plants (n=5). FAC elicitation significantly increased transcripts of: NaLOX3 (b), NaAOS (c), NaAOX (d) and NaCYP94B3-like1/2. Transcripts levels were measured at 1h after FAC-elicitation. Due to high sequence similarity between NaCYP94B3-like1 and NaCYP94B3-like2, qPCR primers we used were not able to distinguish these two copies. g–k, the VIGS-NaLRRK1 plants have enhanced JA biosynthesis and metabolism. FAC elicitation induces significantly higher levels of OPDA (g), OH-JA-ile (i) and COOH-JA-ile (k) in VIGS-EV plants than EV plants, but only marginally higher levels of JA (b) and JA-ile (j) (n=7). l, the VIGS-NaLRRK1 plants accumulated higher transcript levels of the transcription factor NaMyb8 than did VIGS-EV plants. m-o, the VIGS-NaLRRK1 plants accumulated higher transcript levels for the defense genes NaTD2 (m) and NaTPI (n) and higher levels of TPI activity (o) than did VIGS-EV plants. (p) M. sexta gained significantly less mass when fed on VIGS-NaLRRK1 plants than on VIGS-EV plants (n=24). The wounding + FAC treated leaf samples were collected at 1 hr after the treatment for gene expression analysis and at 24 hr after the treatment for TPI activity analysis. In all panels, data are presented as means ± SEM. Asterisk indicates significant difference (*, p<0.05; **, <0.01, *** p<0.001, Student’s t test) between wounding + FAC treatment and control (wounding + water).

DOI: 10.7554/eLife.19531.015
The following figure supplements are available for figure 7:

Figure supplement 1. NaLRRK1 transcript abundance was successfully reduced in VIGS-NaLRRK1 plants in comparison to controls.
DOI: 10.7554/eLife.19531.016

Figure supplement 2. VIGS-NaLRRK1 plants accumulated higher levels of FAC-induced soluble sugars and invertase activity in comparison to control.
DOI: 10.7554/eLife.19531.017
compared to EV (Figure 7g). Consistently, the transcript levels of genes involved in OPDA biosynthesis, such as NaLOX3 and NaAOS, were all significantly increased in VIGS-NalRRK1 plants in comparison to EV plants (Figure 7 a–d), suggesting that NalRRK1 negatively regulates OPDA biosynthesis. Interestingly, the levels of JA and JA-Ile were not significantly different (Figure 7 h and i). However, both the levels of hydroxylated JA-Ile (12OH-JA-Ile) and transcripts of NaCYP94B3-like1/2 - the homologue of AtCYP94B3 that mediates hydroxylation of JA-Ile in N. attenuata (Luo et al., 2016) - were significantly increased in FAC-induced VIGS-NalRRK1 plants in comparison to VIGS-EV plants (Figure 7 e, f, j and k). Since reduced expression of NaCYP94B3-like1/2 results in lower levels of 12OH-JA-Ile and higher levels of JA-Ile (Luo et al., 2016), it is likely that the increased NaCYP94B3-like1/2 transcript accumulations enhanced the hydroxylation of JA-Ile. These results suggest that NalRRK1 negatively regulates both JA biosynthesis and the hydroxylation of JA-Ile, and potentially suppress the effects the defense responses elicited by JA-signaling.

The examination of the downstream JA-dependent defensive traits in VIGS-NalRRK1 plants revealed that the net effect was a negative regulation of JA-signaling. In N. attenuata, the transcription factor NaMYB8, whose expression is activated by increased levels of endogenous jasmonates (Kaur et al., 2010), upregulates the expression of NaTD2 and NaTPI, two key anti-herbivore defensive enzymes (Kang et al., 2006b). In comparison to VIGS-EV plants, VIGS-NalRRK1 plants accumulated significantly higher transcript levels of FAC-induced NaMYB8, NaTD2 and NaTPI (Figure 7 i–o), consistent with the observed effects on JA signaling. In addition to NaMYB8 regulated genes, FAC induced JA signaling also induces changes in primary metabolism, in particular soluble sugars and invertases activity in N. attenuata (Machado et al., 2015). Here, we also found that VIGS-NalRRK1 plants showed higher levels of induced soluble sugars and invertases activity in comparison to those of VIGS-EV plants (Figure 7—figure supplement 2). Consistently, these higher levels of defensive responses in VIGS-NalRRK1 plants resulted in lower growth rates of M. sexta larvae in comparison to those feeding on VIGS-EV plants (Figure 7 p).

Taken together, the data suggest that the HAE-induced M4 gene NalRRK1 and jasmonate signaling form negative feedback loops that regulate and fine tune the induced defenses in N. attenuata.

Whole genome duplications shaped the evolution of HAE-induced EDS network in Nicotiana

Having identified the M4 module, we were interested in exploring the evolution of HAE-induced EDS networks in Nicotiana by analyzing the evolutionary history of genes in the M4 module. For this, we first analyzed the most recent duplication event for each N. attenuata gene using the species reconciliation approach (Materials and Methods). Among all M4 module genes, 79.5% were retained in the genome of Nicotiana and Solanum after at least one round of duplication since the divergence of eudicots from monocots; this percentage retention is significantly higher than the genome-wide average (odds ratio = 1.96, p=2.2E-16, exact binomial test, Table 2). This suggests that gene duplications played a significant role in the evolution of the HAE-induced EDS network. Because Solanaceae taxa experienced a whole genome triplication (WGT) event (Sato et al., 2012), we compared the contributions of the Solanaceae WGT event with Nicotiana-specific lineage duplications (LD) to the evolution of the M4 module genes. A majority of the most recent duplication events of the M4 module genes occurred in the Solanaceae branch (51.5%), likely due to the WGT (Sato et al., 2012). This percentage is significantly higher than the genome-wide average (odds ratio = 1.46, p=1.79E-10, exact binomial test, Table 2) (Figure 8). Because our phylogenomic approach can not specifically distinguish the ancient segmental duplications from the WGT events, we further identified a subset of genes that is located in the syntenic blocks that resulted from the Solanaceae WGT. These genes are consistently significantly enriched in the EDS network (odds ratio = 1.40, p=2.4E-7). In contrast, only ~8.0% of the genes originated from Nicotiana lineage species duplications, which is not different from the genome-wide level (odds ratio=0.84, p=0.12, exact binomial test, Table 2). In addition, when considering genes that were significantly induced by FAC in all three species, N. attenuata, N. acuminata and N. linearis (‘conserved EDS’), similar patterns were found (Table 2). These results suggest that Solanaceae WGT contributed more than lineage specific duplication events to the evolution of the HAE-induced EDS network.

Preferential gene retention followed by genome-wide duplications has been suggested as one of the major mechanisms for network evolution and expansion. Because the complete EDS network before the Solanaceae WGT is unknown, it is difficult to directly test the preferential gene retention
Table 2. Genes from multiple copy gene families and genes containing DTT-NIC1 TE insertions within 1kb upstream region are significantly enriched in the Nicotiana EDS network. The total number of genes used to test gene duplications and the DTT-NIC1 insertions analyses differed due to the additional filtering processes used in the former analysis. For the gene duplication analysis, we excluded all genes whose most recent duplication event was uncertain. WGT: whole genome triplication; NLD refers to Nicotiana lineage specific duplications; complete EDS refers to all of genes identified in the M4 module; conserved EDS refers to M4 genes that were significantly induced by FAC in all three species, N. attenuata, N. acuminata and N. linearis; genome-wide patterns were calculated based on all of genes that were expressed in Nicotiana leaves (normalized FPKM greater than 5 in at least three samples). Bold font color highlights the statistically significant values. Odds ratios were calculated by the following formula: Odd = (p1/(1 – p1))/(p2/(1 – p2)), where p1 is the percentage of genes that are part of the EDS network among testing group, e.g., genes from multiple gene families or genes retained from Solanaceae WGT, and p2 is the percentage of genes that are part of EDS network among all leaf expressed genes.

<table>
<thead>
<tr>
<th>Genome wide</th>
<th># genes from multiple copy families</th>
<th># genes retained from Solanaceae WGT</th>
<th># genes retained NLD</th>
<th>Total number of genes after filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete EDS</td>
<td>9691</td>
<td>6181</td>
<td>1249</td>
<td>14,642</td>
</tr>
<tr>
<td>Conserved EDS</td>
<td>906</td>
<td>587</td>
<td>87</td>
<td>1140</td>
</tr>
<tr>
<td>Odds ratio</td>
<td>1.97</td>
<td>1.45</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>p value</td>
<td>&lt; 2.2E-16</td>
<td>3.30E-10</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>Complete EDS</td>
<td>561</td>
<td>355</td>
<td>65</td>
<td>692</td>
</tr>
<tr>
<td>Conserved EDS</td>
<td>2.18</td>
<td>1.45</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>Odds ratio</td>
<td>&lt; 2.2E-16</td>
<td>1.71E-06</td>
<td>0.41</td>
<td></td>
</tr>
</tbody>
</table>

DOI: 10.7554/eLife.19531.018

hypothesis. Instead, we examined a prediction that would result from the preferential gene retention scenario: for a given gene pair, the observed number of gene pairs that are both found in the EDS is higher than the number of two genes found in the EDS by chance. To test this prediction, we identified a subset of gene pairs (4292 pairs including 8584 genes) that resulted from the Solanaceae WGT (duplications on the Solanaceae branch) in which at least two of the three original copies were retained in the N. attenuata genome and expressed in leaves. Because both members of these gene pairs were both retained in the genome, we specifically examined the process of preferential gene recruitment to the M4 module. Among this subset of genes, 428 were found in module M4, and of those, both members of 120 gene pairs were in the M4 module. This is significantly higher than the expected number from the null model which assumes independent recruitment of two duplicated copies in the M4 module (p<2.2E-16, χ² test), consistent with the prediction that duplicated genes were preferentially retained in the M4 module. Annotations of 120 gene pairs showed that more than 30.8% were either transcription factors or protein kinases, the proportion of which was significantly higher than by chance (2.56% among 4292 pairs, p<2.2E-16, binomial test).

Consistent with the genome-wide analysis, most of the M4 module genes that were previously functionally characterized in N. attenuata, including NaHER1, NaCDPK4/5, NaMEK2, NaER1 and NaJAR4/6, NaACO1/2 evolved from the Solanaceae WGT event (Figure 8), and their homologs were also known to be involved in EDS in Arabidopsis (Dong et al., 2004; Jung et al., 2007; Ludwig et al., 2004; Mewis et al., 2005; Romeis and Herde, 2014; Staswick and Tiryaki, 2004) suggesting that their ancestral copies were likely already involved in EDS. Among these genes, the gene pairs of NaHER1/2, NaCDPK4/5, NaJAR4/6 and NaACO1/2 were highly co-expressed and all members were retained in the M4 module.

Taken together, these results showed that gene duplications, and likely preferential gene retention followed by WGT, shaped the evolution of Nicotiana HAE-induced EDS networks.

Discussion

Regulation of herbivore induced defenses requires a complex and fine-tuned network (Bonaventure et al., 2011; Erb et al., 2012; Wu and Baldwin, 2010), as its fitness cost/benefit ratio
Figure 8. Solanaceae WGT contributed to the evolution of HAE-induced defense signaling in Nicotiana. (a) the gene duplication history of Nicotiana attenuata genes after the divergence of eudicots and monocots. Phylogenetic tree constructed based on one-to-one orthologue genes. The bars under each branch depict the percentage of duplications that occurred at a given branch. The green bar indicates the genome-wide (all genes expressed in leaves) pattern; red bar indicates the duplication events found in module M4. (b) WGT contributed to the evolution of genes that are involved in HAE-induced EDS. All genes are shown as ellipses, and phytohormones as circles. The color of every ellipse shows the most recent duplication events for each gene: blue and gray indicate the Solanaceae WGT, and ancient (shared with Arabidopsis) duplications, respectively. The dots underneath each gene represent the number of homologues found in the N. attenuata genome, and the color indicates whether the homologue was induced by FAC in N. attenuata. Red: significantly induced (FDR adjusted p<0.05, fold change greater than 1.5), black: not induced. Dashed lines indicate the indirect functional interactions. TD: THREONINE DEAMINASE; HGL-DTGs: 17-hydroxysteroidinolases; TP: trypsin proteinase inhibitor. (c–e) the co-expression patterns between the two homologous genes that likely resulted from the Solanaceae WGT; The expression values of each gene were from control and FAC-induced samples from the six Nicotiana species. (f–h) phylogenetic trees showing the duplication history of NaCDPK4/5 (f), NaHER1/2 (g) and NaACO1/2 (h). The blue dot on the phylogenetic tree indicates the duplication events shared among Solanaceae species. The node colors indicate which species the homolog sequences belong to: At (black): Arabidopsis thaliana; Vt (turquoise): Vitis vinifera; Pt (light blue): Populus trichocarpa; Na (yellow): N. attenuata; SI (green): Solanum lycopersicum. DOI: 10.7554/eLife.19531.019

depends on both the type of herbivore attacking the plant and the environmental context of the attack (Baldwin and Hamilton, 2000; Glawe et al., 2003; Heil and Baldwin, 2002; Baldwin, 1998). Therefore HAE-induced early defenses signaling that allows a plant to distinguish herbivore attack from wounding plays an important role in this process (Bonaventure et al., 2011; Howe and Jander, 2008; Wu and Baldwin, 2010). However, identifying herbivore induced EDS is challenging, due to its specificity among different tissues, time points, treatments and overall complexity (Howe and Jander, 2008; Wu and Baldwin, 2010). In this study, we took a novel comparative transcriptomic and co-expression network analysis approach using the leaf transcriptome data from six closely related species that were treated with different HAE, which resulted in the identification of a co-expression network that represents the HAE-induced EDS in Nicotiana. This approach assumes that
if two genes are functionally connected (co-expressed), the expression changes of one gene will also affect the other one during evolution, thus increasing the statistical power for detecting co-expressed gene modules. Although this assumption cannot be applied to species specific co-expression modules, it can be used to identify gene co-expression modules that are conserved among the studied species (Table 1), which likely are functionally important.

Using comparative transcriptomic and network analysis, we identified a co-expressed gene network (module M4), in which 782 genes represent the conserved HAE-induced EDS in Nicotiana. Large numbers of transcription factors and protein kinases were found in this network, suggesting rapid transcriptional and post-transcriptional regulations induced by HAE, which then likely led to the re-configuration of whole-plant metabolism to allow for the production of defense responses (Gulati et al., 2013). Interestingly, among these 782 genes in the HAE-induced EDS, of which only 28.2% and 11.7% in N. obtusifolia and N. mierii were elicited by FAC, respectively, at least 67.6% were induced by O3, in each of these two species (Figure 4). These results suggest that the signaling network, while not elicited by FAC, remains intact in these two species, likely due to changes in FAC-perception. These results are also consistent with the analysis of module conservation which suggested that the M4 module is highly conserved among different Nicotiana species (Table 1).

Molecular signaling cascades often involve negative and positive feedback loops and form circuits. The M4 module is always co-activated and likely upstream or parallel to the activation of JA signaling, indicating that M4 module and JA signaling are likely involved in such circuits. We found NaLRRK1, a FAC-induced hub gene from the M4 module, and jasmonate signaling form negative feedback loops and are co-activated by HAE elicitation among different species that diverged several millions of years ago. The conserved co-activation between jasmonate signaling and NaLRRK1 by HAE and its negative effect on insect performance suggests that the identified feedback loops are important for plant fitness in Nicotiana. Although increased expression of NaLRRK1 after HAE elicitation may lower anti-herbivore defenses, it may increase the net fitness by reducing fitness costs associated with induced jasmonate signaling, such as the changes in primary metabolites that are important for a plant’s tolerance of tissue removal and regrowth (Machado et al., 2013b). Clearly more components are involved in the NaLRRK1 and jasmonate signaling negative feedback loops, such as transcription factors and other protein kinases, which are also likely present in the M4 modules. Future molecular studies that identify direct interacting components with NaLRRK1 will shed light on the mechanisms of the NaLRRK1-JA feedback loops. The challenge will be to quantitatively analyze both transcriptional and post-transcriptional regulation of candidate genes at different time points after elicitation, since their interactions might be highly specific.

Analyzing the gene duplication history of the genes in the M4 module suggested that preferential gene retention after the WGT shared among Nicotiana spp. and Solanum spp. likely have contributed to the evolution of HAE-induced EDS in Nicotiana. We found that more than 30.8% of duplicated pairs that resulted from the Solanaceae WGT, of which both copies were retained in the M4 module, are either transcription factors or protein kinases. This proportion is significantly higher than by chance (p=2.2E-16), consistent with the dosage compensation hypothesis, which predicts that dosage-sensitive genes, of which transcription factors and protein kinases are examples, are more likely to be retained in the signaling network after genome-wide duplications (Edger and Pires, 2009; Freeling, 2009; Hakes et al., 2007; Maere et al., 2005; Wapinski et al., 2007).

Duplicated genes retained in the same network were often considered as evidence of functional redundancy (De Smet and Van de Peer, 2012; Veron et al., 2007); however, genetic redundancy is often evolutionarily unstable and is unlikely to be maintained over long timescales (De Smet and Van de Peer, 2012). The Solanaceae WGT event can be dated to 91–92 million years ago (Sato et al., 2012), yet many of the duplicated gene pairs remain co-expressed after HAE-elicitation. This suggests that retaining these duplicated copies in the same network has been beneficial to plants, likely as a result of increased network complexity and robustness (De Smet and Van de Peer, 2012). This is consistent with the results of studies on examining the function of NaCDPK4/5 and NaJAR4/6 by simultaneously silencing either member of both copies in N. attenuata (Kang et al., 2006; Wang et al., 2007; Yang et al., 2012). When NaCDPK4 and NaCDPK5 were individually silenced, HAE-induced JA accumulations were not affected; silencing both copies increased JA accumulation upwards of 3-fold, which resulted in significant negative fitness effects when plants were not attacked (Yang et al., 2012). Therefore, retaining both NaCDPK4 and NaCDPK5 in the network may increase the robustness against the negative fitness effects resulted...
from null mutations in the gene itself or its regulatory systems (Gu et al., 2003). Interestingly, the duplicated gene pair, NaJAR4/6, which catalyzes the formation of JA-ile, showed additive effects on HAE-induced JA-ile accumulations, since silencing each individual copy both resulted in reduced JA-ile levels (Stasswick et al., 2002) and reductions in the activation of downstream defense responses. Thus, retaining both copies in the network resulted in higher level of HAE-induced JA-ile and defense responses.

In addition to gene duplications, expansions of transposable elements (TEs) can also contribute to the evolution of induced signaling networks. For example, in rice, mPing, a miniature inverted-repeat transposable elements (MITE) family, rapidly expanded in specific strains and its insertions into the 5' flanking region rendered adjacent genes inducible by abiotic stresses by introducing cis-regulatory elements and/or epigenetic markers (Naito et al., 2009; Yang et al., 2005; Yasuda et al., 2013). Similarly, we observed that insertions of the DTT-NIC1, a Solanaceae specific MITEs family that contains a stress inducible cis-regulatory element, the W-box, into 5' regulatory regions of genes are significantly enriched among genes of the HAE-induced EDS network in Nicotiana (p=0.00049, Appendix 1). We speculate that the genome-wide expansions of DTT-NIC1 may have facilitated gene recruitment into the Nicotiana EDS network by introducing cis-regulatory elements into the 5' flanking regions of Nicotiana genes. However, several other mechanisms could also result in the same observation; for example, biotic stresses may mobilize TEs, which are more likely to insert into genes with open chromatin under stressed conditions. Future studies that measure the contribution of DTT-NIC1 insertions in the inducibility of the identified EDS genes by manipulating the DTT-NIC1 insertions sites are needed to examine these hypotheses.

Materials and methods

Sample collection and RNA-sequencing

Plant material was collected as previously reported (Xu et al., 2015). In brief, the seeds of six Nicotiana species were germinated and grown in a York chamber under a 16/8 hr light/dark, 26°C and 65% relative humidity regime until the rosette stage. Manduca sexta and Spodoptera littoralis oral secretions (OS) were collected on ice from larvae reared on N. attenuata plants until the 3rd-5th instar as previously described (Halitschke et al., 2001). To simulate herbivore attack, one leaf of each plant was wounded with a pattern wheel and 20 μL of 1:5 diluted OSsix, OSs or FAC (138 ng μL−1 C18:3-Glu) or water (as control) was added to the puncture wounds. All leaves were collected 30 min after elicitation, their mid-veins rapidly excised, flash frozen in liquid nitrogen and stored at −80°C until analysis. For each species and treatment, three biological replicates were used based on common practice of RNA-seq experiments. The phytohormone data for all samples were analyzed and published in (Xu et al., 2015).

Total RNA was extracted from ~100 mg aliquots of homogenized leaves that were used for phytohormone analysis (Xu et al., 2015) using Trizol (Thermo Fisher Scientific, Germany) according to the manufacturer’s protocol. All RNA samples were subsequently treated with RNase-free DNase-I (Thermo Fisher Scientific) to remove all genomic DNA contamination. The mRNA was enriched using the mRNA-seq sample preparation kit (Illumina), and ~200 bp insertion size libraries were constructed using the Illumina whole transcriptome analysis Kit following the manufacturer’s standard protocol (Illumina, HiSeq system). All libraries were then sequenced on the Illumina HiSeq 2000 at the sequencing core facility at the Max Planck Institute for Molecular Genetics, Berlin. On average, more than 35 million 50-nt paired-end raw reads for each sample were obtained. All raw reads are deposited in the NCBI short reads archive (SRA) under the project number PRJNA301787.

Gene co-expression network and differential expression analysis

All raw sequence reads were trimmed using AdapterRemoval (v1.1) (Lindgren et al., 2012) with parameters ‘-collapse -trim3 -trimquality 2 -minlength 36’ before being used for transcriptome assembly. We mapped all reads to the N. attenuata genome (release v2) using TopHat2 (v2.0.6) (Trapnell et al., 2009). We used parameter ‘-segment-mismatches 2 -read-gap-length 2 -m 0 -N 2’ for N. attenuata RNA-seq reads and allowed more mismatches for the other five species with parameters ‘-segment-mismatches 3 -read-gap-length 5 -m 1 -N 7’. The mapping statistics are shown in Supplementary file 1C. The reads count matrix was then extracted from the bam files using HTSeq
with parameters `-a 1 -t exon` (Anders et al., 2015). For both mapping and reads counting, N. attenuata gene models were used. We further simulated the sequence divergence and estimated the expression levels to evaluate the effects of sequence divergence on reads mapping and gene expression estimation. The analysis revealed that sequence divergence did not affect the overall gene expression estimations using our mapping protocol.

We constructed the co-expression modules using the R package weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) based on the trimmed mean of M-values (TMM) normalized FPKM (Fragments per kilobase of transcript per million mapped reads) values. Because the expression values among samples clustered by species, we applied a parametric normalization to reduce the effects that resulted from the background expression differences among species using the ‘Combat’ function from the sva package (Johnson et al., 2007). The normalized wounded and FAC-induced samples were first used to calculate the soft connectivity and the top 5000 connected genes were selected for module construction. The power selection, module significan- cence and intra-module connectivity analysis were performed according to the WGCNA tutorial (http://goopy.gl/tw20K). We only selected the genes in module M4 with membership greater than 0.75 for downstream analysis. FAC-induced hub genes were defined as the top 5% most connected genes in FAC-induced transcriptomes. In total, 67 genes were selected as hub genes. For visualization, the M4 module was exported to Cytoscape (3.1.0) with the edge weight greater than 0.15 as a cutoff.

Preservation of the M4 module at co-expression levels among four different species was analyzed using the ‘modulePreservation’ function from WGCNA. All pairs-wise comparisons were performed based on RNA-seq data from samples treated with WW, OSw0, FAC and OSp0. The M4 module genes that were not expressed in a given species were excluded. In total, 99.4% of genes were used for the module preservation tests.

The ratio of Ka/Ks (ω) was calculated using one-to-one orthologue genes between N. attenuata and N. obtusifolia using KaKs calculator (Zhang et al., 2006) with ‘YN’ method. Because the calculations of ω is unreliable for gene pairs with extremely low Ks values, all gene pairs with Ks value less than 0.02 were excluded. In total 1182 genes pairs (88.9%) were used for the analysis.

We identified differentially expressed genes using the edgeR (Robinson et al., 2010) package based on the raw count data. Genes with greater than 1.5-fold change and FDR-adjusted p-values less than 0.05 were considered as differentially expressed. For both the gene co-expression network and differential expression analyses, we only considered genes that had a FPKM greater than 5 in at least three samples. The venn diagram analyses for differentially expressed genes among species were performed using the R package vennuler (Wilkinson, 2011).

**Gene duplication detection**

To identify gene duplication events, we first assigned homologous groups (HG) using a similarity-based method. To do so, we used all genes that were predicted from 11 eudicot genomes (Xu et al, submitted). In brief, all-vs-all BLASTP was used to compare the sequence similarity of all protein coding genes, and the results were filtered based on the following criteria: E-value less than 1E-20; match length greater than 60 amino acids; sequence coverage greater than 60% and identity greater than 50%. All BLASTP results that remained after filtering were clustered into HGs using the Markov cluster algorithm (mcl) (Enright et al., 2002).

For each of the identified HGs, we constructed a phylogenetic tree using an in-house developed pipeline. In brief, we aligned all coding sequences for each HG using MUSCLE (v.3.8.31) (Edgar, 2004) based on translated protein sequences with TranslatorX (v.1.1) (Abascal et al., 2010). For all aligned sequences, all non-informational sites (gaps in more than 20% of sequences) were removed using trimAL (v.1.4) (Capella-Gutierrez et al., 2009). Then, for each HG, PhyML (v. 20140206) (Guindon et al., 2009) was used to construct the gene tree with the best nucleotide substitution model estimated based on jModeltest2 (v.2.1.5) (Darriba et al., 2012) with the following parameters: -f i -g 4 -s 3 -AIC -a. The support for each branch was calculated using the approximate Bayes method (PhyML).

Duplication events within each HG were predicted based on the reconstructed gene trees using a tree reconciliation algorithm, which compares the structure of species and gene trees to infer duplication events (Page and Charleston, 1997). This approach allows one to predict the history of gene duplication events at each branch of the species tree. To reduce the false positives, we only
considered tree structures with approximated Bayes support greater than 0.9 at all three closest branches for assignment of gene duplication events.

**Microarray hybridization and data analysis**

We measured the FAC-induced gene expression changes in WT and JA-deficient plants (ir-AOC) using microarray analyses. The WT samples were the same as the samples used for the RNA-seq analysis. The germination and growth conditions, FAC elicitation, sample collection and RNA extraction for the analysis of the ir-AOC plants were the same as those described for the analysis of the WT plants. cDNA preparation and hybridizations were performed as described in Kallenbach et al. (Kallenbach et al., 2011). Quantile normalization and log2 transformation was performed for all raw microarray data using the R package ‘AffyLM4PreProcess’ (http://goo.gl/TJnA6Q). Probes with 1.5-fold change and adjusted p-values less than 0.05 were considered differentially expressed. The sequences of all probes were mapped to the N. attenuata draft genome (v1.0), and only the probes that uniquely mapped to annotated gene regions were considered for downstream analysis. All microarray data were deposited in the public GEO (Gene Expression Omnibus) repository (GSE75006).

**Functional annotation of genes**

The gene functional annotation process was part of the N. attenuata genome sequencing effort (Xu et al, submitted). Multiple annotation tools were used. In brief, BLAST2GO (Gotz et al., 2008) was used to annotate the GO terms for all predicted genes, and the GO enrichment analysis was performed using the ClueGO (v2.1.1) (Bindea et al., 2009) plugin in Cytoscape. In addition, all N. attenuata genes were annotated using MapMan (Thimm et al., 2004) with annotation information from Arabidopsis, tomato, potato and cultivated tobacco. The transcription factors and protein kinase containing genes were identified based on the identified domains in each gene according to the criteria described in Pérez-Rodríguez et al. (Riano-Pachon et al., 2007) using the ITAK tool (http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi). All N. attenuata genes from NCBI were retrieved and compared to the predicted N. attenuata genes using BLAST and the best hits were annotated accordingly. The functional annotation of all N. attenuata genes and N. attenuata genome data are available from the N. attenuata database server (http://nadh.ice.mpg.de/NaDH/). The R scripts used for this study and original data are available as source code file and source data.

**qPCR confirmation of gene expression**

Total RNA was extracted from ∼50 mg leaves using Trizol (Thermo Fisher Scientific) according to the manufacturer’s protocol. In brief, all RNA samples were subsequently treated with RNase-free DNAse I (Fermentas) to remove all genomic DNA contamination. All cDNA samples were synthesized from ∼1 µg total RNA using SuperScript II reverse transcriptase (Thermo Fisher Scientific). The relative transcript accumulation levels of selected genes were measured using qPCR on a Stratagene MX3005P PCR cycler (Stratagene). For all qPCRs, the elongation factor-1A gene (NaEF1a, accession number: D63396) was used as the internal standard for normalization as previously described (Johnson et al., 2007). The primer pairs for qPCR are listed in Supplementary file 1D. All qPCR reactions were performed using qPCR core kit for SYBR Green I (Eurogentec) in a 20 µL reaction system. At least four biological replicates were used for all qPCR measurements.

**Characterizing the regulation of NaLRRK1 expression**

To characterize the expression of NaLRRK1, WT plants and three transformed lines were used: ir-AOC (line A-457) (Kallenbach et al., 2012) ir-COI (line A-249) (Paschold et al., 2007) and 35S-mj/ir-mj (line A-204) (Stitz et al., 2011). Seed germination procedure was the same as described above. Seedlings were transferred to Teku pots ten days after germination and then were planted into 1L pots in the greenhouse, which was maintained at 26–28°C under 16 hr of light as described in (Krügel et al., 2002).

The FAC elicitations were the same as described above. For the flg22 treatment (W+flg22), 20 µL of 100 nM flg22 in water was immediately applied to standardized puncture wounds produced by the fabric pattern wheel. To test the effects of JA-ile on the expression of NaLRRK1, 0.25 µM or 0.125 µM JA-ile in 20 µL FAC (containing 12.5% ethanol), was immediately applied to the puncture
wounds in leaves. The leaf samples (n=5) were collected 1 hr after treatment. All leaf samples were flash frozen in liquid nitrogen, and stored at −80°C until analyzed.

**Construction of reporter fusions and subcellular localization**

The construction of NaLRK1-YFP reporter fusion was carried out as described by Earley et al. and Ran et al. (Earley et al., 2006; Li et al., 2014), and a reporter fusion was also constructed for the A. thaliana plasma membrane (PM) intrinsic protein 2a (accession number: X75883) which was previously characterized as a marker for membrane associations (Shibata et al., 2016). The open reading frame (ORF) of NaLRK1 and PM were firstly amplified with Phusion Green High-Fidelity DNA polymerase (Thermo) by primer pairs listed in the *Supplementary file 1D;* an additional sequence (CACC) was then introduced into the forward primers to facilitate directional cloning into the pENTR/D-TOPO vector (Thermo). The reconstructed plasmids were transformed into E. coli TOP10 competent cells, then amplified and isolated as the ‘entry vector’ for the Gateway cloning. The ‘entry vector’ containing the ORF of NaLRK1 or PM was recombined into destination vectors using LR clonase (Invitrogen) to form a C-terminal NaLRK1-YFP and C-terminal PM-CFP. Recombined plasmids were transformed into E. coli TOP10 competent cells, and then transformed into A. tumefaciens strain GV3101 for subsequent plant transformation. The transformation was performed using A. tumefaciens strain GV3101 following the protocol by Green et al. (Green et al., 2012). Fluorescence was visualized 48 hr following the inoculation with a Zeiss LSM 510 Meta confocal microscope (Carl Zeiss, Jena, Germany). The images were analyzed using LSM 2.5 image analysis software (Carl Zeiss, Inc.).

**Virus induced gene silencing (VIGS) of candidate genes expression and phytohormone measurements**

VIGS based on the tobacco rattle virus (TRV) was used to transiently knock down the expression of the candidate gene NaLRK1 in N. attenuata as previously described (Galis et al., 2013). In brief, fragments of ~300 bp of target genes were amplified by PCR with primers listed in *Supplementary file 1D.* PCR fragments were recovered by agarose gel electrophoresis and purified using a gel band purification kit (Amersham Biosciences) according to the manufacturer’s instructions, and subsequently digested with BamHI and SalI and inserted into plasmid, pTV00 (RNA1). After sequencing to validate the constructs, pTV-fragment-VIGS constructs and pTV00 (empty vector), together with RNA2, were transformed into *Agrobacterium* for the VIGS procedure.

At 21 days after *Agrobacterium* inoculation, rosette-stage plants were wounded with a pattern wheel and 20 µL of 1:5 diluted FAC (138 ng µL⁻¹ C18:3-O-glucosidase before dilution) or water was added to the puncture wounds. All samples were collected at 1 hr after elicitation with mid-veins excised, flash frozen in liquid nitrogen, and stored at −80°C until analysis. Silencing efficiency was quantified by qPCR. Overall, more than 88% of the target transcripts were silenced by VIGS.

Phytohormones were analyzed as described previously (Wang et al., 2007). In brief, ~100 µg frozen leaf was homogenized in a Genogrinder with 0.8 mL ethylacetate spiked with [9,10,2-H]<sub>3</sub>-dihydro-JA and [13C<sub>3</sub>]-JA-ile. Homogenates were centrifuged for 30 min at 4°C and the organic phase was collected and evaporated to dryness, which were subsequently reconstituted in 300 mL of 70% (v/v) methanol/water for analysis on an advance UPLC (Bruker), equipped with column ZORBAX eclipse XDB (Agilent) and quantified on an EVOQ triple quadrupole mass spectrometer (Bruker) using the MRM transitions described in (Schäffer et al., 2016).

**Secondary metabolites analysis and bioassay in VIGS plants**

To quantify secondary metabolites that were known to function defensively in N. attenuata, leaves of VIGS-EV and VIGS-NaLRK1 (n=8) plants were treated with W+FAC for 24 hr, harvested and ground in liquid nitrogen and stored at −80°C until analysis.

Trypsin proteinase inhibitor (TPI) assay was carried out as previously described (van Dam et al., 2001). Briefly, 100 µg of ground powder (n=6) was extracted in a protein extraction buffer. The protein content was determined using the Bradford method and PI activity was analyzed with the radial diffusion assay, using soybean trypsin inhibitor (STI) as the external standard.

Soluble sugars (glucose, fructose and sucrose) and starch concentrations were quantified as described by Machado et al. (Machado et al., 2013a). Briefly, soluble sugars were extracted from
plant tissue (n=6) using 80% (v/v) ethanol, followed by an incubation step (20 min at 80°C). The precipitate was collected by centrifugation (15 min, 11,000 g, 4°C). Pellets were re-extracted twice with 50% (v/v) ethanol. Supernatants from all extraction steps were pooled and enzymatically quantified for sucrose, glucose and fructose. The remaining pellets were used for an enzymatic determination of starch.

To evaluate the performance of the specialist herbivore _M. sexta_ on transformed plants, neonates were allowed to feed on EV and transformed plants (n=28) and their masses were measured at 0, 6, 10 and 14 d after transfer to experimental plants. To ensure that all larvae were at a similar developmental stage and had similar body mass at the start of the bioassay, newly hatched neonates were placed on untreated WT leaves for 48 hr and weighed. The neonates with similar size were selected for the bioassays.

35S-JMT/IR-ME: _N. attenuata_ transgenic plants ectopically expressing _Arabidopsis_ (_Arabidopsis thaliana_) jasmonic acid O-methyltransferase (35S-jmt) and with _N. attenuata_ methyl jasmonate esterase silenced with RNAi.

Acknowledgements

We thank T Krügel and the glasshouse team of the MPICOE for taking care of plants, P Langfelder for suggestions on data analysis, S Pottinger for help with sample collections, VIGS experiments and measure gene expression with qPCR, M Schäfer and M Kallenbach for setting up methods for the phytohormone analysis, R Li for helping with the cellular localization experiments and K Gase for helping with the RNA-seq data submission to NCBI. We thank J Wu, C Hettenhausen, M Huber and PM Schlüter for constructive comments on the manuscript and E Gaquerel for design suggestions of the Figure 8b. This work was supported by Swiss National Science Foundation (No: PEBZP3-142886 to SX), the Marie Curie Intra-European Fellowship (IEF) (No: 328935 to SX), Alfred and Anneliese Sutter-Stöttner foundation (to SX), the Max Planck Society and the European Research Council advanced grant ClockworkGreen (No. 293926 to ITB).

Additional information

Competing interests

ITB: Senior editor, _eLife_. The other authors declare that no competing interests exist.

Funding

<table>
<thead>
<tr>
<th>Funder</th>
<th>Grant reference number</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung</td>
<td>PEBZP3-142886</td>
<td>Shuqing Xu</td>
</tr>
<tr>
<td>European Research Council</td>
<td>293926</td>
<td>Ian T Baldwin</td>
</tr>
<tr>
<td>European Commission</td>
<td>328935</td>
<td>Shuqing Xu</td>
</tr>
<tr>
<td>Max-Planck-Gesellschaft</td>
<td></td>
<td>Wenwu Zhou, Thomas Brodmüller, Zhihao Ling, Ashton Omdahl, Ian T Baldwin, Shuqing Xu</td>
</tr>
<tr>
<td>Sutter-Stötter-Stiftung</td>
<td></td>
<td>Shuqing Xu</td>
</tr>
</tbody>
</table>

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

WZ, Acquisition of data, Analysis and interpretation of data; TB, Constructed phylogenetic trees and identified all gene duplications events for each gene family, Analysis and interpretation of data; ZL, Performed RNA-seq mapping and quantification and assisted the RNA-seq data analysis, Analysis and interpretation of data; AO, Performed motif analysis, Analysis and interpretation of data; ITB,
Additional files

Supplementary files

- Supplementary file 1. Six supplementary tables. (A) Genes induced by FAC in all six Nicotiana species. Green and red colors indicate e3 ubiquitin-protein ligase and transcription factors, respectively. (B) Genes which showed lower expression in samples induced by M. sexta OS than by FAC in both N. attenuata and N. pseudacanthorum. Numbers in each column indicate the TMM-normalized FPKM values. (C) Summary of RNA-seq reads from six closely related Nicotiana species. (D) Primers used for qPCR and ViPS in this study.

DOI: 10.7554/eLife.19531.020

Major datasets

The following datasets were generated:

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
<th>Dataset title</th>
<th>Dataset URL</th>
<th>Database, license, and accessibility information</th>
</tr>
</thead>
</table>

The following previously published dataset was used:

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
<th>Dataset title</th>
<th>Dataset URL</th>
<th>Database, license, and accessibility information</th>
</tr>
</thead>
</table>

References


Zhou et al. eLife 2016;5:e19531. DOI: 10.7554/eLife.19531


Zhou et al. eLife 2016;5:e19531. DOI: 10.7554/eLife.19531

25 of 30

153


Appendix 1

Enrichment of the DTT-NIC1 in the regulatory region of the Nicotiana EDS genes

Results
In addition to gene duplications, we further analyzed the putative role of TE insertions on the evolution of EDS in Nicotiana. Our previous work had revealed that several miniature inverted-repeat transposable elements (MITE) families, including DTT-NIC1, a Solanaceae-specific subgroup of the Tc1/Mariner family, expanded in Nicotiana (Xu et al, submitted). Permutation tests showed that insertions sites of annotated MITE families are enriched more than four times at 1kb upstream of genes compared to random chance (p<0.001, permutation test). To examine the role of MITE insertions on the evolution of the HAE-induced EDS, we compared the probability of genes recruited to the EDS that contain MITE insertions within 1 kb upstream to the genome-wide pattern. The results showed that MITE insertions into the 1 kb upstream regions significantly increased the probability of genes being recruited into the M4 module (odds ratio = 1.3, p= .0E-3, exact binomial test, Appendix 1—table 1).

Appendix 1—table 1. Insertions of the MITEs family DTT-NIC1 is significantly enriched in 1kb upstream region of M4 module genes. p-values were calculated based on exact binomial test. Superfamilies are represented using different letters in the names: DTT for Tc1/Mariner, DTM for Mutator, DTA for hAT, DTH for PIF/Harbinger. Statistically significant one is highlighted in red.

<table>
<thead>
<tr>
<th>MITEs name</th>
<th>Number of insertions to regulatory region of M4 module genes</th>
<th>Odd ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTA-NIC1</td>
<td>4</td>
<td>0.54</td>
<td>0.32</td>
</tr>
<tr>
<td>DTA-NIC2</td>
<td>3</td>
<td>1.14</td>
<td>0.75</td>
</tr>
<tr>
<td>DTA-NIC3</td>
<td>22</td>
<td>1.38</td>
<td>0.16</td>
</tr>
<tr>
<td>DTA-NIC4</td>
<td>3</td>
<td>0.55</td>
<td>0.49</td>
</tr>
<tr>
<td>DTA-NIC5</td>
<td>1</td>
<td>0.23</td>
<td>0.19</td>
</tr>
<tr>
<td>DTA-NIC6</td>
<td>4</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>DTA-NIC7</td>
<td>1</td>
<td>0.72</td>
<td>1.00</td>
</tr>
<tr>
<td>DTA-NIC8</td>
<td>1</td>
<td>0.61</td>
<td>1.00</td>
</tr>
<tr>
<td>DTA-NIC9</td>
<td>12</td>
<td>1.42</td>
<td>0.28</td>
</tr>
<tr>
<td>DTH-NIC1</td>
<td>23</td>
<td>1.53</td>
<td>0.063</td>
</tr>
<tr>
<td>DTH-NIC2</td>
<td>23</td>
<td>1.14</td>
<td>0.56</td>
</tr>
<tr>
<td>DTH-NIC3</td>
<td>11</td>
<td>1.28</td>
<td>0.38</td>
</tr>
<tr>
<td>DTT-NIC1</td>
<td>65</td>
<td>1.63</td>
<td>0.0049</td>
</tr>
</tbody>
</table>

DOI: 10.7554/eLife.19531.021

Analyses on the insertion of each individual MITE family separately revealed that the increased probability of EDS recruitment was mainly due to DTT-NIC1 to 1kb upstream region of genes (Appendix 1—table 2 and Supplementary file 1C). In contrast, DTT-NIC1 insertions into downstream regions did not increase the likelihood of genes being recruited to the EDS (odds ratio = 1.1, p=0.5). Analyses of the genes that were significantly induced by FAC elicitation in all three FAC-responding species, N. attenuata, N. acuminata and N. linearis, showed similar patterns but had an even higher odd ratio (Appendix 1—table 1). These data suggest that insertions of DTT-NIC1 into gene regulatory regions significantly contributed to the recruitment of genes to the Nicotiana HAE-induced EDS.
### Appendix 1—table 2: Seven DNA motifs are significantly enriched in the 1kb upstream region of M4 module genes. The presence of the motif in sequences and p-values were obtained using the HOMER package. Motif sequences were annotated by searching the plant cis-regulatory sequence database in PlantPan2 (http://plantpan2.ips.ncu.edu.tw/).

<table>
<thead>
<tr>
<th>Motif sequence</th>
<th>Annotation</th>
<th>Presence in M4 module genes (%)</th>
<th>Presence in background dataset (%)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGTCAACG</td>
<td>W-box1</td>
<td>60.5</td>
<td>38.5</td>
<td>1.0E-53</td>
</tr>
<tr>
<td>TTGACCWT</td>
<td>W-box2</td>
<td>69.0</td>
<td>48.4</td>
<td>1.0E-47</td>
</tr>
<tr>
<td>NNACGCGT</td>
<td>unknown</td>
<td>19.7</td>
<td>10.0</td>
<td>1.0E-23</td>
</tr>
<tr>
<td>TTCACAGG</td>
<td>unknown</td>
<td>18.7</td>
<td>11.6</td>
<td>1.0E-12</td>
</tr>
<tr>
<td>GTCGGCTG</td>
<td>unknown</td>
<td>39.5</td>
<td>29.7</td>
<td>1.0E-12</td>
</tr>
<tr>
<td>CGAAGACT</td>
<td>unknown</td>
<td>32.0</td>
<td>23.3</td>
<td>1.0E-11</td>
</tr>
<tr>
<td>TTTCCAGG</td>
<td>unknown</td>
<td>34.3</td>
<td>25.9</td>
<td>1.0E-10</td>
</tr>
</tbody>
</table>

DOI: 10.7554/eLife.19531.002

One of the mechanisms by which DTT-NIC1 insertions could contribute to the recruitment of a gene to the HAE-induced EDS is the introduction of promoter binding motifs. To examine this potential mechanism, we performed de novo identifications of enriched transcription factor binding motifs from regulatory regions among HAE-induced EDS genes. Within the 1kb upstream region sequences, seven significantly enriched motifs (p<1E-10) were found among the M4 module genes in comparison to the genome-wide background (Supplementary file 1D). Among them, two W-box motifs (W-box1: AGTCAACG and W-box2: TTGACCWT) were the most statistically significant. These conserved motifs are known to be bound by WRKY transcription factors that are induced by biotic stresses (Chen et al., 2012; Gao et al., 2016; Zou et al., 2011). W-box1 and W-box2 box were found in the 1kb upstream regions of more than 60.5% and 69.0% of M4 module genes, respectively. A similar pattern was also found for the EDS genes that were induced by FAC in all three FAC-responding species.

Interestingly, at the genome-wide level, the distribution of the W-box1 motif in the N. attenuata genome is significantly associated with the insertions of DTT-NIC1 (p<0.001, based on 1000 permutation test). Analyzing the consensus sequence of the DTT-NIC1 in N. attenuata also showed that DTT-NIC1 contains the W-box1 motif. Further comparisons of insertion sites of DTT-NIC1 and motif locations in the 1kb upstream region of N. attenuata HAE-induced EDS genes revealed that at least 44.4% of the DTT-NIC1 insertions introduced W-box1 to the regulatory region of adjacent genes. These results suggest that DTT-NIC1 mediated W-box insertions into the regulatory region of genes might have contributed to the recruitment of genes to the HAE-induced EDS network in Nicotiana.

### Method

We performed de novo motif identification using HOMER (Heinz et al., 2010), which identifies motifs that are significantly enriched in the test dataset in comparison to background datasets. The genes from the M4 module and leaf expressed genes that were not involved in the M4 module were used as positive and negative datasets, respectively. To reduce false positives, we performed additional analysis by identifying enriched motifs from 100 randomly selected subsets of genes from the negative dataset. These were then compared with the enriched motifs in the M4 module. We considered all of the enriched motifs from the M4...
module that occurred in more than 5% of motifs identified from the random selected dataset as noise and removed them from downstream analysis. Using this approach, we identified seven 8 bp motifs (Appendix 1—table 2) that are significantly enriched in the regulatory region of M4 genes (p<1e-10). The presence and absence of the identified W-box motif in the N. attenuata genome were analyzed using the `scanMotifGenomeWide` function from the HOMER package (Heinz et al., 2010).

Insertions sites of DTT-NIC1 to the regulatory region of genes were identified as described in (Xu et al, submitted). In brief, using all of the MITE consensus sequences as a library, we searched the 1kb upstream region of genes with RepeatMasker. The parameters `--nomax -nois -cutoff 250` were used. The enrichment analysis of MITEs in 1kb upstream region of M4 module genes and associations between the MITEs and identified motifs were analyzed using permutation tests. For this, we randomly shuffled the insertion sites of MITEs 1000 times using the `shuffle` function from the BEDTOOLS package (Quinlan and Hall, 2010). We then calculated the null models of MITEs insertion sites and overlaps between MITEs insertions and motif locations in the N. attenuata genome. These null models were then used to determine the enrichment significance of observed insertions of MITEs families in the 1 kb upstream region of M4 module genes as well as observed associations between locations of MITEs families and motifs.
6 Manuscript III
“Nicotiana attenuata Data Hub (NaDH): an integrative platform for exploring genomic, transcriptomic and metabolomic data in wild tobacco”

(Brockmöller T, Ling Z, Li D, Gaquerel E, Baldwin IT, Xu S)
**DATABASE**

* Nicotiana attenuata Data Hub (NaDH): an integrative platform for exploring genomic, transcriptomic and metabolomic data in wild tobacco

Thomas Brockmoller, Zhihao Ling, Dapeng Li, Emmanuel Gaquerel, Ian T. Baldwin and Shuqing Xu

**Abstract**

**Background:** *Nicotiana attenuata* (coyote tobacco) is an ecological model for studying plant-environment interactions and plant gene function under real-world conditions. During the last decade, large amounts of genomic, transcriptomic and metabolomic data have been generated with this plant which has provided new insights into how native plants interact with herbivores, pollinators and microbes. However, an integrative and open access platform that allows for the efficient mining of these omics data remained unavailable until now.

**Description:** We present the *Nicotiana attenuata* Data Hub (NaDH) as a centralized platform for integrating and visualizing genomic, phylogenomic, transcriptomic and metabolomic data in *N. attenuata*. The NaDH currently hosts collections of predicted protein coding sequences of 11 plant species, including two recently sequenced *Nicotiana* species, and their functional annotations, 222 microarray datasets from 10 different experiments, a transcriptomic atlas based on 20 RNA-seq expression profiles and a metabolomic atlas based on 895 metabolite spectra analyzed by mass spectrometry. We implemented several visualization tools, including a modified version of the Electronic Fluorescent Pictograph (eFP) browser, co-expression networks and the Interactive Tree Of Life (ITOL) for studying gene expression divergence among duplicated homologous. In addition, the NaDH allows researchers to query phylogenetic trees of 16,305 gene families and provides tools for analyzing their evolutionary history. Furthermore, we also implemented tools to identify co-expressed genes and metabolites, which can be used for predicting the functions of genes. Using the transcription factor *NaMYB8* as an example, we illustrate that the tools and data in NaDH can facilitate identification of candidate genes involved in the biosynthesis of specialized metabolites.

**Conclusion:** The NaDH provides interactive visualization and data analysis tools that integrate the expression and evolutionary history of genes in *Nicotiana*, which can facilitate rapid gene discovery and comparative genomic analysis. Because *N. attenuata* shares many genome-wide features with other *Nicotiana* species including cultivated tobacco, and hence NaDH can be a resource for exploring the function and evolution of genes in *Nicotiana* species in general. The NaDH can be accessed at: http://nadhi.ce.mpgd.de/.

**Keywords:** *Nicotiana attenuata*, Phylogenomics, Transcriptomics, genomics, Metabolomics, Co-expression analysis

* Correspondence: swalice.mpgd.de
1 Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Hans-Scharre-Straße 8, D-07745 Jena, Germany
Full list of author information is available at the end of the article

© The Author(s), 2017 Open Access. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

*Nicotiana attenuata*, is a diploid wild tobacco native to the Great Basin desert of the United States with populations across Utah, Nevada, Arizona, Oregon, and California. This plant has adapted to an ecological niche defined by the post-fire environment, where soils tend to be nitrogen-rich and biotic stresses are highly dynamic [1]. During the last decade, *N. attenuata* has been developed as a model organism to study plant-environment interactions in its native environment [2–6], and a large number of transcriptomic and metabolomic datasets have been generated with this plant. For example, more than 230 transcriptomic data from *N. attenuata* have been submitted to the NCBI GEO database. However, to efficiently analyze, explore and visualize such genome-wide metabolomic and transcriptomic data remain challenging for individual researchers. In particular, most of these data were not centralized and integrated. Recently, we sequenced and annotated the genomes of *N. attenuata* and its close relative *N. obtusifolia* [7], which provided an opportunity to create tools for centralizing, integrating and visualizing these omics data from this plant.

Specialized metabolites are of special importance in the defenses of plants, therefore, understanding their regulation and their evolutionary history are of central interests in plant biology. However, identifying genes involved in the biosynthesis of specialized metabolites remains difficult due to the large number of gene duplication events in plant genomes, and the structural diversity of the metabolites produced by plants. Recently, studies suggest that co-expression analysis is a powerful tool to rapidly identify genes involved in the biosynthesis of specialized metabolites, because many of these genes are often co-expressed [8]. However, such co-expression analysis often involves large amounts of data and remains difficult to handle for researchers who are not familiar with sophisticated statistics and lack programming skills.

Here, we present the *Nicotiana attenuata* Data Hub (NaDH, http://nadh.ice.mpq.de), a centralized publicly available platform for storing and integrating genomic, transcriptomic and metabolomic data from *N. attenuata* (Fig. 1). To provide user-friendly data analysis and visualization, we implemented tools from the Electronic Fluorescent Pictograph (eFP) browser, co-expression networks and the Interactive Tree Of Life (iTOL). Using the genes from the biosynthetic pathway of phenolamides as an example, we show that NaDH users can rapidly identify genes involved in specialized metabolites and make inferences on their evolution history.

Construction and content

Genomic data

The NaDH includes 33,449 and 27,911 predicted gene models from *N. attenuata* (release r2.0) and *N. obtusifolia* genomes (release r1.0), respectively. For comparative genomic analysis, additional gene sequences and structures from nine dicot plant genomes are also included in the database (Table 1). To provide functional annotations, the predicted enzyme commission (EC) identities, gene ontology (GO) terms, and protein domains are included.

![Fig. 1](image-url) Overview of data structure and utilities in the NaDH. The NaDH consists of 10 major utilities, which can be accessed by either gene ID or metabolite spectrum ID.
Table 1 Overview of included species for comparative genomics

<table>
<thead>
<tr>
<th>Species</th>
<th>Version</th>
<th># of genes</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. attenuata</td>
<td>v2.0</td>
<td>33,449</td>
<td><a href="http://naiti.ice.mpg.de/NaiDH/download/">\text{http://naiti.ice.mpg.de/NaiDH/download/}</a></td>
<td>[7]</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>TAIR 10</td>
<td>27,416</td>
<td><a href="http://phytozone.jgi.doe.gov/arabidopsis">\text{http://phytozone.jgi.doe.gov/arabidopsis}</a></td>
<td>[9, 10]</td>
</tr>
<tr>
<td>C. sativus</td>
<td>v1.0</td>
<td>21,503</td>
<td><a href="http://phytozone.jgi.doe.gov/cucumber">\text{http://phytozone.jgi.doe.gov/cucumber}</a></td>
<td>[12]</td>
</tr>
<tr>
<td>M. guzunus</td>
<td>v2.0</td>
<td>28,140</td>
<td><a href="http://phytozone.jgi.doe.gov/mimusus">\text{http://phytozone.jgi.doe.gov/mimusus}</a></td>
<td>[13]</td>
</tr>
<tr>
<td>P. trichocarpa</td>
<td>v3.0</td>
<td>41,335</td>
<td><a href="http://phytozone.jgi.doe.gov/poplar">\text{http://phytozone.jgi.doe.gov/poplar}</a></td>
<td>[14]</td>
</tr>
<tr>
<td>S. lycoptum</td>
<td>ITAG2.3</td>
<td>34,727</td>
<td><a href="http://phytozone.jgi.doe.gov/tomato">\text{http://phytozone.jgi.doe.gov/tomato}</a></td>
<td>[15]</td>
</tr>
<tr>
<td>S. tuberosum</td>
<td>v3.4</td>
<td>35,119</td>
<td><a href="http://phytozone.jgi.doe.gov/potato">\text{http://phytozone.jgi.doe.gov/potato}</a></td>
<td>[17]</td>
</tr>
<tr>
<td>V. vinifera</td>
<td>GenomeScope 12X</td>
<td>26,346</td>
<td><a href="http://phytozone.jgi.doe.gov/grape">\text{http://phytozone.jgi.doe.gov/grape}</a></td>
<td>[18]</td>
</tr>
</tbody>
</table>

Gene families and phylogeny

The NaDH includes 23,340 homologous groups constructed based on protein coding sequences from 11 eudicot species (Table 1). PhyML was used to construct phylogenetic trees with high confidence from these homologous groups that contain more than two genes. In total, 16,305 trees containing 255,404 genes (of which 28,610 are from *N. attenuata*) are included in the NaDH. In addition, 81,459 gene duplication events detected from high confidence phylogenetic trees (approximate Bayes branch supports of greater than 0.9 for the target node and its two child nodes) using the species-overlapping algorithm implemented in Notung-2.6 [19, 20] are also included in the database. The majority of gene duplication events in *N. attenuata* are found at the Solanaceae branch (Table 2), consistent with the observation that species of Solanaceae share a whole-genome triplication event.

Transcriptomic data

The NaDH contains expression profiles from both RNA-seq and microarray datasets [21–23]. For the RNA-seq datasets (Illumina HiSeq 2000, pair-end sequencing, NCBI accession number: PRJNA317743), the expression level (transcript per million, TPM) [24] of each gene from different tissues sampled from leaves, seeds, roots, stems and flowers are included (Table 3). In total, 21,970 genes were expressed in at least one tissue (TPM greater than 5). Roots contain the largest number of expressed genes (Table 3). For the microarray dataset, 222 microarrays (based on Agilent platform: GPL13527) from *N. attenuata* leaves, roots and flowers are included. The probes of this microarray platform were mapped to the *N. attenuata* genome and the uniquely mapped probes were annotated according to gene predictions. In total, this microarray platform contains the expression profiles of 27,374 predicted *N. attenuata* genes. The microarray datasets are organized according to their corresponding experiments and the detailed information on the genotypes, developmental stages, treatments of the plants that provided the samples (Additional file 1) are provided.

Metabolomic data

Metabolomic data from 14 isolated tissues of *N. attenuata* growing under controlled conditions in glasshouse were curated. This includes a pool of all non-senescing rosette leaves, combined lower, middle and higher segments of the stem, the complete root system, dried seeds, complete floral buds of 8 mm length, complete sepal ring, the nectary, the ovary (not including the nectary), the style, anthers, filaments (not including anthers), and the corolla tube and limb, collected at anthesis. Pools of 100 mg tissues were extracted using 80% methanol. Independent extractions were also conducted with 20% methanol. Samples were analyzed using UHPLC-ESI/qTOF-MS in positive ion mode. MS/MS data collection was achieved via a previously-described pipeline [25] and 895 reconstructed MS/MS spectra were obtained [26]. This MS/ MS dataset has also been deposited in the EMBL EBI open metabolomics database MetaboLights: www.ebi.ac.uk (accession no. MTBLS335).

Table 2 The number of detected duplication events in *N. attenuata*

<table>
<thead>
<tr>
<th>Duplication time</th>
<th># of duplications</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>N. attenuata</em> specific</td>
<td>3,929</td>
</tr>
<tr>
<td>Shared among <em>Nicotiana</em> spp.</td>
<td>2,577</td>
</tr>
<tr>
<td>Shared among Solanaceae</td>
<td>6,760</td>
</tr>
<tr>
<td>Shared with M. guzmunus</td>
<td>240</td>
</tr>
<tr>
<td>Shared among core eudicots</td>
<td>8,548</td>
</tr>
<tr>
<td>Total</td>
<td>22,054</td>
</tr>
</tbody>
</table>

A large number of duplication events is shared among all Solanaceae, which supports a shared whole-genome triplication event.
6 Manuscript III

Table 3 Overview of RNA-Seq data

<table>
<thead>
<tr>
<th>Library ID/ SRA ID</th>
<th>Tissue</th>
<th>Treatment/development stage</th>
<th># of expressed genes</th>
<th>Additional note on sampling procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA14989R</td>
<td>Root</td>
<td>Rosette stage plants, treated with 5 µL 1:1 diluted M. sexta oral secretion three times in leaves</td>
<td>15,499</td>
<td>Roots of rosette stage plants that were treated three times on leaves were collected for RNA extraction. The treatments were performed at 10 am and 6 pm on the day before sampling and 10 am on the day of sampling. Samples were collected at 11 am.</td>
</tr>
<tr>
<td>SRX1084095</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA15000LE</td>
<td>Leaf</td>
<td>Rosette stage plants, treated with 5 µL 1:1 diluted M. sexta oral secretion three times in leaves</td>
<td>12,179</td>
<td>Local leaves of rosette stage plants that were treated three times on leaves were collected for RNA extraction. The treatments were performed at 10 am, 6 pm on the day before sampling and 10 am on the day of sampling. Samples were collected at 11 am.</td>
</tr>
<tr>
<td>SRX1084996</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA17171X</td>
<td>Leaf</td>
<td>Rosette stage plants, no treatment</td>
<td>11,840</td>
<td>Rosette stage leaves were collected for RNA extraction. Samples were collected at 11 am.</td>
</tr>
<tr>
<td>SRX1084554</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA15045TT</td>
<td>Stem</td>
<td>Rosette stage plants, treated with 5 µL 1:1 diluted M. sexta oral secretion three times in leaves</td>
<td>14,682</td>
<td>Stems of rosette stage plants that were treated three times on leaves were collected for RNA extraction. The treatments were performed at 10 am, 6 pm on the day before sampling and 10 am on the day of sampling. Samples were collected at 11 am.</td>
</tr>
<tr>
<td>SRX1084090</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1605CO</td>
<td>Corolla</td>
<td>Early developmental stage, no treatment</td>
<td>13,662</td>
<td>Samples were collected in the afternoon, 60 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1615CO</td>
<td>Corolla</td>
<td>Late developmental stage, no treatment</td>
<td>13,486</td>
<td>Samples were collected at 6 pm (open flower) and at 9 am (closed flower after opening), 4-10 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1506STI</td>
<td>Stigma</td>
<td>Mature stigma, no treatment</td>
<td>14,485</td>
<td>Stigma samples were collected in the afternoon, 40 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1507POL</td>
<td>Pollen tube</td>
<td>No treatment</td>
<td>3,490</td>
<td>Pollen tubes were pooled.</td>
</tr>
<tr>
<td>SRX1084003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA15085NP</td>
<td>Style</td>
<td>Mature style without pollination</td>
<td>13,492</td>
<td>Styles were collected at 7 am, anthers were removed one day before, and 50 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1509STO</td>
<td>Style</td>
<td>Mature style, pollinated with pollens from different genotype</td>
<td>13,365</td>
<td>Styles were collected at two hours after pollination, at 7 am. Anthers were removed one day before, and 30 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA15105TS</td>
<td>Style</td>
<td>Mature style, self-pollinated</td>
<td>13,333</td>
<td>Styles were collected at two hours after pollination, at 7 am. Anthers were removed one day before, and 30 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1511NEC</td>
<td>Nectary</td>
<td>Mature nectary, no treatment</td>
<td>12,928</td>
<td>Samples were collected in the afternoon, 60 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1512ANTE</td>
<td>Anther</td>
<td>Mature anther no treatment</td>
<td>11,550</td>
<td>Samples were collected in the afternoon, 60 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1515COV</td>
<td>Ovary</td>
<td>Mature ovary, no treatment</td>
<td>13,960</td>
<td>Samples were collected in the afternoon, 60 samples were pooled.</td>
</tr>
<tr>
<td>SRX1084009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1516PRD</td>
<td>Pedicel</td>
<td>Mature pedicel, no treatment</td>
<td>14,550</td>
<td>Samples collected at 9 am (heading down) and 4 pm (heading up) were pooled.</td>
</tr>
<tr>
<td>SRX1084010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1516DRL</td>
<td>Flower</td>
<td>Fully opened flowers, no treatment</td>
<td>14,390</td>
<td>Both morning (7 am) and evening (6 pm) flowers were collected, 1 sample of each were pooled.</td>
</tr>
<tr>
<td>SRX1084011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1517FIL</td>
<td>Flower</td>
<td>Two early developmental stages of flowers, no treatment</td>
<td>16,434</td>
<td>Samples were collected at 6 pm, 1 bud and 1 middle stage flower were collected. Sepals were removed from the samples.</td>
</tr>
<tr>
<td>SRX1084012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA150018E</td>
<td>Seed</td>
<td>Treated with liquid smoke</td>
<td>9,227</td>
<td>100 ng seeds treated with 1:50 diluted liquid smoke solution for 9-15 min were used for RNA extraction.</td>
</tr>
<tr>
<td>SRX1084087</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1503SVL</td>
<td>Seed</td>
<td>Treated with water</td>
<td>8,872</td>
<td>100 ng seeds treated with water for 9-15 min were used for total RNA extraction.</td>
</tr>
<tr>
<td>SRX1084098</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1503SEED</td>
<td>Seed</td>
<td>Dry seeds</td>
<td>8,681</td>
<td>100 ng dried seeds directly used for total RNA extraction.</td>
</tr>
<tr>
<td>SRX1084099</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The raw reads information and methods used for generating these data are available under NCBI accession number PRJNA317743.

Gene-to-gene co-expression
To facilitate the identification of co-regulated genes in *N. attenuata*, we calculated the pairwise expression correlation co-efficiency based on RNA-seq data from 20 different tissues using three different methods: Gini, Spearman and Pearson [27]. In total, 15,216 informative genes (with TPM greater than 5 in at least one tissue and a variance greater than 1) were used and gene pairs with absolute expression similarity greater than 0.65 were considered for the final dataset.
All data are stored in NaDH and can be visualized in a network graph.

**Metabolite-to-metabolite co-expression**

Metabolite-to-metabolite tissue associations were calculated using Ochiai similarity based on binary metabolite dataset (containing 14 tissue vectors), with a cutoff of 2 on the ZMAD transformed values [28]. The metabolite-to-metabolite pairwise associations were calculated across the dataset by comparing each metabolite spectrum with the other metabolite spectra [26]. All data are stored in the underlying database for fast accessing under the utilities of NaDH.

**Gene-to-metabolite co-expression**

Gene-to-metabolite tissue associations were calculated using Ochiai similarity with binary gene and metabolite data (presence and absence) generated from 12 shared tissues between the transcriptome dataset and the metabolome dataset (a cutoff of transcriptome dataset is ZMAD transformed TPM greater than 3, and a cutoff of metabolome dataset is ZMAD transformed value greater than 2). Gene expression was centered by median and median-absolute-deviation (MAD) to obtain a relative expression level [29]. In total, 23,075 genes and 895 metabolite spectra with expression levels above the threshold were used for the network constructions [26]. The pairwise correlations were calculated using Ochiai correlations based on the transformed binary values [28] and only Ochiai correlation coefficient (occ) greater than 0.3 were considered for the final dataset. The resulting network is based on a correlation matrix between 18,046 genes and 887 metabolite spectra.

**Metabolite structure similarity**

Metabolite structure similarity was calculated from pairwise MS/MS alignments based on spectral fragment similarity and common neutral losses similarity (NL). A standard normalized dot product (NDP), also referred to as the cosine correlation method for spectral comparison, was applied for the calculations of spectral fragment similarity. The NL-based similarity between individual MS/MS was implemented using a list of 52 neutral losses (NLs) commonly encountered during tandem MS fragmentation as well as more specific ones that had been previously annotated for MS/MS spectra of *N. attenuata* secondary metabolite classes [25].

**Database architecture and implementation details**

An overview of the NaDH database architecture is shown in Fig. 1. All data are connected to the gene and metabolite. The data storage function was implemented using the open-source relational management system MySQL to store all data except the genome and gene sequence information. The website was developed using Kohana - an open source, object-oriented model-view-controller (MVC) web framework (https://kohanaframework.org/). The unique filesystem design of the Kohana web framework allows a modular design of all NaDH features and enables easy implementations of new functions in the future. The visualization of gene and metabolites expression was implemented by a modified version of the eFP Browser [30]. For interactive visualization of the different co-expression networks, we implemented the open-source graph visualization tool Cytoscape.js (https://github.com/cytoscape/cytoscape.js).

**Utility and discussion**

**Search functions**

Search functions are implemented for finding genes and metabolites. Genes of *N. attenuata* can be searched by name, functional annotation (InterPro domain, EC number, Gene Ontology identifier), orthologous genes in other species and sequence similarity based on BLAST. Metabolite spectra and fragments can be searched by name, metabolite class and measurements (m/z value, retention time). In both scenarios, a table summarizes information of the corresponding genes or metabolite spectra, and guides users to downstream analysis.

**Gene-to-gene co-expression analysis**

This function can be used to understand the regulatory mechanisms and predict putative functions of genes. The input is the identifier of a gene, and the outputs are genes co-expressed with the input gene above the user defined threshold (correlation coefficient) and this is presented in both an interactive network graph and table formats. In the resulting co-expression network graph, each node represents a co-expressed gene, radiating position of the node represents the most recent duplication events it has experienced and clock-wise position of the node represents the region with the highest expression among four tissues (leaves, roots, flower buds, and seeds) and relative tissue specificity. The resulting table shows more detailed functional annotation of each node. Figure 2a shows an example output for the transcription factor NaMYB8 (NIAIV7_g41919).

**Metabolite-to-metabolite co-expression analysis**

This function can be used to find co-regulated metabolite spectra, which might indicate co-occurrence in biosynthetic pathways and signal cascades. The input of this function is the identifier of a metabolite spectrum of interest, and all co-expressed metabolite spectra above the user-defined threshold and the results are presented in co-expression network graph and table formats. In the co-expression network, each node represents a metabolite spectrum, the color of
the nodes represents the annotated class of the corresponding metabolite spectrum, and the edge represents the structural similarity between two nodes: a yellow edge for NDP and red edge for NL. The network can be re-arranged based on expression similarity values or annotated metabolite classes.

**Gene-to-metabolite co-expression analysis**

Co-expression between gene and metabolite can be used to both infer putative functions of the genes and to identify candidate biosynthetic pathways of the metabolites [31–34]. In the NaDH, we provide a function to find bi-directional searches for co-expressed genes and metabolite
spectra. For searching metabolite spectra that are co-expressed with a gene of interest, the input is an identifier of the gene and the output is a metabolite spectra network with each node representing a metabolite spectrum and each edge representing the structural similarity between two metabolite spectra. In order to search for genes that co-express with a metabolite spectrum of interest, the input is an identifier of the metabolite spectrum and the output is the gene co-expression network with each node representing a gene and the position of the node representing the duplication history and expression (similar to gene-gene co-expression network graph). Figure 2b shows the co-expressed metabolite spectra for the gene NaAT1 (NIVATv7.g11614).

Phylogenetic analysis
In the NaDH, a phylogenetic tree can be directly uploaded and visualized with ITOL [35, 36]. The input is an identifier of the gene of interest, and the output is a phylogenetic tree that integrates the expression of N. attenuata genes among 20 different tissues. In addition, the intron-exon structures were also integrated with the phylogenetic tree to provide further information on the evolutionary history of the gene. Figure 3c shows an example of the output for the gene DH29 of the phenolamides pathway.

Expression visualization
The expression of genes and metabolites can be visualized via a modified version of the eFP Browser developed by Nicholas J. Provart et al [30]. The input is either the identifier of a gene (or a probe ID from the microarray) or metabolite spectrum of interest and the output is the expression of the gene (or probe) or precursor of a metabolite spectrum mapped to each tissue or treatment, respectively. The expression levels of the gene are shown as a heatmap with yellow and red colors indicate low and high expression, respectively. The binary expression of the precursor of a metabolite spectrum is shown as a heat-map with red and yellow colors indicate expressed and not expressed, respectively. Figure 3a and b show an example output for a gene and metabolite in the eFP Browser, respectively. The expression values are also provided as a table or bar chart for user-specific analysis. In addition, we also implemented a function to compare and visualize the expression of multiple genes and precursors of metabolite spectra among different tissues.

Example analysis
The evolution and diversity of specialized metabolites in plants are largely shaped by gene duplication events [37]. Consequently, to find which of the duplicated copies are involved in the biosynthesis and regulation of specific secondary metabolites is challenging. Using the above-described utilities in the NaDH and genes known to be involved in phenolamides biosynthetic pathway as an example, we show that the integration of gene-to-gene, gene-to-metabolite, metabolite-to-metabolite and gene duplication history can help to identify genes that are involved in specialized metabolites in the genus Nicotiana.

Phenolamides, a group of diverse metabolites abundant in many plant reproductive organs, are rapidly induced after herbivore attack in vegetative tissues of several Solanaceae species and play an important role as induced chemical defenses. The biosynthesis of phenolamides originates from the main phenylpropanoid pathway via N-acyltransferase-dependent conjugation to polyamines or aryl monoamines (Fig. 4) [38, 39]. Similar to the biosynthetic pathways of many other secondary metabolites, genes involved in the phenylpropanoid pathway contain multiple copies (Fig. 4). Because several genes involved in the regulation and biosynthesis of phenolamides have been functionally characterized in N. attenuata, this group of metabolites provides an ideal example to test the utility of the NaDH.

One of the key components that regulates the biosynthesis of phenolamides in N. attenuata is the R2R3-MYB transcription factor, NaMYB8 (NIVATv7.g41919) [40]. We first searched for all genes that were co-expressed with NaMYB8 with a cutoff with a gini correlation coefficient (gcc) greater than 0.7, which resulted in 2,620 co-expressed genes. Among these genes, we searched for homologs that are putatively involved in biosynthetic steps of the main phenylpropanoid pathway. Although in each step, several copies were found in N. attenuata, only one or two copies were co-expressed with NaMYB8. Among them, functional characterization using virus-induced gene silencing (VIGS) revealed that AT1, CV86 and DH29 are indeed involved in the biosynthesis of herbivore-induced phenolamides, such as caffeoylputrescine (CP) and N,N'-dicafeoylserpinidine (DCS), which function as anti-herbivore chemical defenses [40]. The duplication history of these genes also showed that most of the recent duplications of these genes were from the Solanaceae branch, suggesting the whole genome triplication event of the Solanaceae contributed to the evolution of herbivore-induced phenolamides in Nicotiana. Additional co-expression analysis of gene-metabolite and metabolite-metabolite associations showed that the key metabolites (phenylalanine, MCS, CP and DCS) and genes (NaPAL, NaC3H, NaDH29) from the pathway can be retrieved by searching highly co-expressed genes and metabolites (Fig. 4b). Although such co-expression analysis can only be used for the metabolites that are synthesized in the tissues that they accumulate in, these results suggest that using the utilities implemented in the NaDH, users can rapidly identify co-expressed genes and metabolites that are involved in the same pathway.
Fig. 3 Example of visualizing gene expression and evolution. a Expression of NaDX29 (NIATV1_g09682) in leaves and roots of control or wounding or INA4194-induced plants. b Presence and absence of N,N-di-(2-furfuryl)perimidine (DCS) [m/z: 470.2286±2.87, retention time: 3.16±0.16950] in different N. attenuate tissues. c Phylogenetic tree of 4C1 gene family visualized using iTOL. Gene structure of each gene is shown on the right side. Colors refer to different species.
Conclusion

We present the NaDH, which integrates genomic, transcriptomic, and metabolomic data in *N. attenuata* and provides useful tools for the interactive visualization of gene expression divergence and gene duplication history. Additional tools for finding co-expressed genes and metabolites can facilitate rapid gene discovery for specialized metabolites in *N. attenuata* and infer their evolutionary paths. Since the most of genome-wide features are shared among the genus *Nicotiana*, the NaDH can also be used to explore the function and evolution of genes in other *Nicotiana* species.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in NaDH (http://nadh.ice.mpg.de/NaDH/).
Additional file

Additional file 1: Overview of microarray datasets that can be visualized via the eFP browser in NaOH. Different unique experiments are included that consist of different genotypes, treatments and time points. IV: empty vector; WT: wild type; W: induced by water; DS: induced by M. sexta oral secretion; VGS: virus-induced gene silencing; GLV: green leaf volatile; VOC: volatile organic compound. (DOCX 17 kb)

Abbreviations

GLV: Green leaf volatiles; NODP: Normalized dot product; NL: Neutral loss; OS: M. sexta oral secretory; TPM: Transcripts per million; VGS: Virus-induced gene silencing; VOC: Volatile organic compound; WT: Wild type

Acknowledgments

We thank all members of the Molecular Ecology Department, Max Planck Institute for Chemical Ecology for fruitful discussions and valuable comments. We thank Martin Nietzger for providing technical support on setting up the public database and Meret Huber for constructive suggestions on the manuscript.

Funding

The work was supported by Swiss National Science Foundation (project number: PBZPF5-142886 to SIK), the Marie Curie Intra-European Fellowship (EI) (Project Number: 26935 to SIK and the Max Planck Society.

Competing Interests

The authors declare that they have no competing interests.

Authors’ contributions

TB and SV designed and developed the database system, TB, ZL and SV generated gene annotations and analyzed data. EG and OL performed the mass spectrometry analysis, inferred tissue-preferentially of expression for metabolites and developed the gene-tonegene co-expression and metabolite similarity analysis. SV and TB wrote the manuscript. IFL provided the o-cim data originated by the NaOH. All authors have read and approved the final manuscript.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

1Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Hans-Knöll-Str 8, 07745 Jena, Germany. 2Centre for Organismal Studies, Heidelberg University, Imm. Neuenheimer Feld 360, Heidelberg D-69120, Germany.

Received: 30 June 2016 Accepted: 23 December 2016
Published online: 13 January 2017

References

7 Discussion

Understanding the evolution of adaptive traits is challenging, as it requires genomic and transcriptomic information across closely related species, an understanding of the genetic basis of key adaptive traits, as well as sophisticated phylogenomic tools. In this dissertation, I present three manuscripts that provide new insights into the evolution of anti-herbivore defenses in *Nicotiana*. In **manuscript I**, we analyzed the evolution of *Nicotiana* genomes and found that *Nicotiana* experienced a whole genome triplication (WGT) event with other Solanaceae species, as well as a repertoire of recent transposable element (TE) insertions that bloated the genome. Analyzing the evolution of nicotine biosynthetic machinery in detail showed that step-wise duplications of two primary metabolisms lead to the evolution of nicotine biosynthesis, a process further facilitated by TE insertions that likely contributed to the coordinated expression among different nicotine biosynthesis genes. In **manuscript II**, we identified a conserved herbivore-induced early defense signaling (EDS) network in *Nicotiana*, another example for adaptive traits. By analyzing the evolution of each individual gene involved in EDS, we showed that both WGT and insertions of TEs played an important role for the evolution of EDS in *Nicotiana*. In **manuscript III**, we introduced a new database platform that combines different “-omics” data and a comprehensive phylogenomic toolbox that can be used to illustrate the function and evolution of genes as well as the regulation of metabolic pathways in *Nicotiana*.

7.1 The genome assembly of *N. attenuata* as basis for phylogenomic analyses

Genome sequences are of high importance for biological research, as they record the evolutionary history of organisms and the gene sequence information that are both necessary for revealing the function of genes and traits. Thanks to recently developed sequencing technology, the cost of genome sequencing is dropping rapidly (El-Metwally et al., 2013) and many genomes have become publicly available (Chain et al., 2009; Muir et al., 2016). However, due to limitations of sequencing and assembly methodology, it remains a challenge to obtain high quality genome assemblies of species with large genome sizes and genomes that contain high amounts of repetitive
7 Discussion

elements. This is the case for the species used in this dissertation, *N. attenuata*, which has a genome size of 2.5Gb and in which more than 80% of the genome is composed of repetitive sequences (manuscript I).

Prior to our work, five *Nicotiana* genomes were made publicly available, *N. benthamiana*, *N. sylvestris*, *N. tomentosiformis*, *N. otophora* and *N. tabacum* (Bombarely et al., 2012; Sierro et al., 2013; Sierro et al., 2014). Although they do meet “minimum” standards based on sequencing depth and the completeness of their genomic sequences (Chain et al., 2009), they provide little information on the evolution of *Nicotiana* genomes, due to their small scaffold length (see Table 1) and lack of annotations. Furthermore, the highly fragmented *Nicotiana* genomes prevent in-depth analysis on the function of genes, as many genes cannot be correctly annotated, given that different exons are often located on different scaffolds.

Table 1: Genome quality of the five publicly available *Nicotiana* genome including the two newly sequenced species in our study: *N. attenuata* and *N. obtusifolia*. Repeat content and completeness was calculated for all four species in manuscript I.

<table>
<thead>
<tr>
<th></th>
<th><em>N. attenuata</em></th>
<th><em>N. obtusifolia</em></th>
<th><em>N. sylvestris</em></th>
<th><em>N. tomentosiformis</em></th>
<th><em>N. otophora</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size</td>
<td>2.5 Gb</td>
<td>1.5 Gb</td>
<td>2.6 Gb</td>
<td>2.7 Gb</td>
<td>2.7 Gb</td>
</tr>
<tr>
<td>Repeat content</td>
<td>81.00%</td>
<td>64.80%</td>
<td>78.20%</td>
<td>75.30%</td>
<td>NA</td>
</tr>
<tr>
<td>N50 (scaffolds)</td>
<td>524.5 kb</td>
<td>134.1 kb</td>
<td>79.7 kb</td>
<td>82.6 kb</td>
<td>26.6 kb</td>
</tr>
<tr>
<td>completeness</td>
<td>83.87%</td>
<td>85.08%</td>
<td>78.63%</td>
<td>66.13%</td>
<td>NA</td>
</tr>
</tbody>
</table>

To sequence genomes containing a large percentage of repeats and TEs, it is essential to use sequencing techniques that are suitable for long reads in order to increase the possibility of assembling complete repeat and coding regions during post-processing. Short read lengths, such as that produced by Illumina HiSeq with a read length of 100 bp, and which was used for the previous five published *Nicotiana* genomes, are not sufficient for this process (Goodwin et al., 2016). To circumvent this problem, we applied a combined approach by using Illumina short reads, 454 reads and PacBio long reads. Using this approach, we could assemble the *N. attenuata* genome in superb quality with a N50 scaffold length of around 520 kbp (see Table 1) which is similar to tomato or potato but three times the genomic size (Xu et al., 2011; Tomato and Consortium, 2012).
7.1 The genome assembly of N. attenuata as basis for phylogenomic analyses

Using four different methods, including gene resequencing and mapping previously obtained ESTs of N. attenuata and N. benthamiana back to the genome, we determined the high completeness and quality of the N. attenuata genome assembly. Furthermore, TF and gene annotation was performed to build a solid foundation for following analyses, and to this end we manually inspected and curated around 500 annotated genes of biological interest. Among all detected gene models in N. attenuata, 76.3% (25,506 out of 33,449) are expressed in at least one tissues. Additional functional annotation including the identification of GO terms and EC numbers completed the process. By establishing this high quality genome assembly and annotation (as described in manuscript 1), we could provide a solid base for the phylogenomic analysis performed in this thesis, and it will further serve as a reference genome in basic and applied science.

7.2 Gene and genome duplications provide genetic material for the evolution of novel traits

Studies have shown that duplicated genes are one of the major resources that have contributed to the evolution of novel traits (Van de Peer et al., 2009) through the process of neo- and subfunctionalization. In studies among all organisms, a large amount of duplications were detected, especially in the plant kingdom, in addition to many segmental duplications, several polyploidization events were found in angiosperms (Bowers et al., 2003; Tomato and Consortium, 2012). Interestingly, different types of duplications appear to evolve differently and to contribute also to the evolution of adaptive traits differently. Although most duplicated copies are lost followed WGDs, genes encoding multidomain proteins or are dosage-sensitive are often preferentially retained (Gibson and Spring, 1998; Thomas et al., 2006). If the selection acts on both copies to maintain dosage balance, it increases their longevity which might provide additional time for the evolution of novel functions (Teufel et al., 2016). This is in concordance with genome-wide analyses revealing that duplication types are correlated with genetic and metabolic features. While genes involved in transcriptional regulation and signal transduction are often retained after large-scale duplications such as WGDs, enzymes and genes involved in stress responses and secondary metabolism are often retained after local (tandem) gene duplications (Seoighe and Gehring, 2004;
Maere et al., 2005; Hanada et al., 2008; Chae et al., 2014). In manuscript I, we detected a similar pattern and showed that the ERF gene family expanded by several rounds of TD. This gene family contains ERF189 that is an important TF for the activation of the biosynthesis of nicotine, an important secondary metabolite. Although ancient TD events are shared among Petunia, Solanum and Nicotiana, an additional TD happened in each lineage independently, whereas only in Nicotiana did the ERF189 acquire particular root-specific expression patterns within a relative short time frame. A good example from the literature is the class of plant resistant genes (R-genes) that show species-specific amplification through TD in Brassicaceae (Yu et al., 2014). A similar study in Arabidopsis revealed that tandem duplicated regions show high variability among ecotypes, and hypothesized that TD promotes adaptation in rapidly changing environments and are the target of positive selection (Hanada et al., 2008), which might be also the case for the ERF gene family in Solanaceae.

Interestingly, studies show that gene families involved in transcriptional regulation and signal transduction are specifically expanded after WGD events, indicating that this pattern is conserved among all plant species (Blanc and Wolfe, 2004; Rodgers-Melnick et al., 2012). One assumption is that the dosage balance hypothesis drives the functional-related gene retention, since genes involved in signal transduction and transcription show a high dosage effect (Birchler et al., 2001) due to their large involvement in protein complexes where only correct stoichiometrics of the involved elements result in correct function (Birchler et al., 2001). Therefore, we expected to detect a similar effect on the evolution of signaling after herbivore attack in plants. In manuscript II, by analyzing the early defense signaling network that is rapidly activated by the presence of herbivore-associated elicitors in Nicotiana, we observed that a network consisting of 782 co-expressed genes is conserved among six analyzed Nicotiana spp. Among those duplicated copies retained after WGT, a large number of genes belong either to transcriptional factors or protein kinases, suggesting rapid transcriptional and post-transcriptional regulations induced by HAE, which then likely led to the re-configuration of whole-plant metabolism to allow for the production of defense responses, and supports the functional-related gene retention after WGD events.

The duplicated copies can evolve new functions by (i) changing the coding sequence or (ii) changing gene expression and therefore its regulation. Different studies
7.2 Gene and genome duplications provide genetic material for the evolution of novel traits

have shown that the mode of duplication can influence the sequence and expression divergence in different ways. However, studies also indicate that the duplication type has an influence on the following sequence and expression divergence.

In 2004, a study in Arabidopsis showed that 57% of gene pairs retained after WGD and 73% of pairs retained from older duplication diverged in expression, whereas 21% of younger duplicated genes showed asymmetrical evolution in protein sequences (Blanc and Wolfe, 2004). A similar study in rice also showed that the majority of genes duplicated during WGD had been lost or had been sub- or neofunctionalized (Salse J. et al., 2009). Other studies support these findings and indicate that the majority of duplicated genes formed by paleopolyploidy have diverged in gene expression (Wang et al., 2012). However, by comparing different modes of duplication, the most diverged expression can be detected in TE-mediated duplicated genes, whereas WGD and TD derived gene copies show the lowest expression divergence. This finding suggests that novel functions to adapt to new niches and new environments are mainly derived by TE. Consistently, in manuscript 1, we found that the majority of duplicated genes resulting from WGT showed diverged expression patterns, likely through neo- or subfunctionalization. Expression divergence can result in regulatory changes and rewiring of biosynthetic pathways and signaling, which helps organisms to adapt to new niches by dealing with biotic and abiotic stresses. In summary, gene duplications are important to evolve novel functions and represent one of the major motors that shape plant evolution. The mode of duplication, gene structures, and functions are important factors that affect the evolution of duplicated genes in plants.

Of all duplication types, polyploidy events have the highest number of newly introduced genes. Although they have the potential to provide evolutionary and ecological advantages for organisms, as polyploids seem to better survive under detrimental conditions and are adapted to a broad range of habitats, it is controversial whether polyploidy events are evolutionarily advantageous in all circumstances (Snyder and Cheson, 2000; Soltis and Soltis, 2000). Current studies propose that recent polyploids have a higher extinctions rate compared to their diploid relatives (Arrigo and Barker, 2012). It is argued that only successful polyploids survived and overcame genetic and evolutionary difficulties, such as inefficient selection when alleles are
7 Discussion

masked by multiple copies (Meyers and Levin, 2006), meiotic difficulties inherent to duplicated genomes (Ramsey and Schemske, 2002; Brownfield and Köhler, 2011), loss of self-incompatibility (Levin, 1983; Robertson et al., 2011) and, possibly, low effective population size (Lynch and Conery, 2003). However, in periods of environmental change, larger numbers of polyploids arise. One example is the K-T mass extinction event happened around 65 Mya where a large fraction of life disappeared, because of considerable changes of the environment of earth. Interestingly, most polyploidy events happened in the same time frame, including the WGD in the Solanaceae lineage. By analyzing several angiosperms, Fawcett et al. hypothesized that polyploid species had a selective advantage in comparison to their diploid relatives during this period (Fawcett et al., 2009), which might also have been the case for Solanaceae lineage.

In summary, the retention of paralogous genes is important for the evolution of novel functions, and therefore both small-scale and large-scale duplications are an important factor driving this evolution in plants. However, the factors controlling gene retention are still unknown but gene structures and functions as well as the mode of duplication seems to be influential.

7.3 Transposable elements promote rapid expression divergence

TEs were often considered “junk,” and were left out of the genome analyses for long time. However, recent studies, including our work presented in manuscript I and II, show that TEs are another important motor that drives the evolution of the genome and of genes through insertions that affect the evolution of gene regulatory networks and the control of gene expression (Sundaram et al., 2014; Cao et al., 2016). Three mechanisms are implicated in the literature to be involved.

First, epigenetic silencing of TEs affects the expression of nearby genes. Studies show that TE silencing by DNA methylation can reduce expression of closely located genes, whereas this effect is reduced with larger distances to the nearest TEs (Hollister and Gaut, 2009). The observed genome-wide patterns indicate that genes containing TE insertions in their up- and downstream region are generally lower expressed than genes without insertions (Lippman et al., 2004; Hollister and Gaut, 2009). However, the magnitude of this effect differs among species, probably due to different silencing efficiencies in host organisms, among TE-families, and between tissues. Even between
closely related species such as *A. thaliana* and *A. lyrata*, differences are detectable, as distances greater than \(~2.5\) kbp and \(~1.0\) kbp are needed to not affect gene expression, respectively (Hollister et al., 2011). It has been proposed that TE-induced gene silencing might have a deleterious effect on the corresponding genes (Hollister and Gaut, 2009). However, the underlying mechanisms of TE-induced gene silencing are not fully understood.

Second, TEs can carry regulatory elements such as TF-binding sites. This theory is not new, having been already proposed in the 1950s by Barbara McClintock and further supported by several studies (McClintock, 1950; McClintock, 1956). A previous study in humans and mice showed that a high number (about 20%) of all analyzed binding sites are embedded in TEs. However, this number varies significantly between different TF-binding sites, and is cell-type specific (Sundaram et al., 2014). A comparable study is still lacking in plants. However, we detected a significant number of TF-binding sites that were likely derived from TE insertions, indicating that results could be similar. In particular, we showed that around one third of the GCC- and G-box elements, which are known bindings sites for the TFs involved in nicotine regulation (*MYC2* and *ERF189*), are located in the promoter regions of nicotine biosynthesis-involved genes, and may be mediated by TE insertions. These findings are consistent with another study that performed a genome-wide analysis in tomato and potato and observed that a significantly high number of GCC- and G-box sites are located in repetitive elements (Mehra et al., 2015). Interestingly, at the genome-wide level, TE-introduced TF binding sites are different in tomato and potato. While both species are prone to other TF-binding sites; Ibox, SORLIP2, MADS and G-box motifs are highly abundant in potato, whereas L1-box, Hexamer, ACE and GCC motifs are common in tomato (Mehra et al., 2015). This might be the result of species-specific TE activity in the genome.

Third, transcriptional activity of TEs in the genome might provide smRNA targeting sites that result in tissue-specific expression regulation. Studies have revealed a large number of expressed repetitive elements that are involved in small RNA and small interference RNA biogenesis. These RNAs can influence gene expression by *cis*- or *trans*-regulation, post-transcriptional gene silencing, and DNA methylation in plants.
7 Discussion

(Feijo et al., 2009; Malone and Hannon, 2009; Momose et al., 2010; Matzke and Mosher, 2014).

Many TEs are transcriptionally active and have a high transposition rate, and so the effects of TE insertions on gene expression can spread rapidly among a large number of genes and pathways in the genome. Studies in plant species revealed that the genome size of angiosperms varies by more than 1000-fold due to differing TE activity (Tenaillon et al., 2010). The TE content can invade and expand a genome within a short time. For example, maize and *Gossypium* spp. have doubled their genome sizes within only ~5 million years, a process that was mainly facilitated by the expansion of LTRs (SanMiguel et al., 1998; Hawkins et al., 2006). Similarly, as presented in manuscript I, species of the Solanaceae lineages also experienced bursts of TEs in a relative short timeframe, which resulted in genomes containing 81% and 64.8% of TEs in *N. attenuata* and *N. obtusifolia*, respectively. Due to this high rate of transposition, reaching $10^3$ to $10^5$ per element per generation (Biémont and Vieira, 2006; Saha et al., 2008), TEs should have a significant impact on the widespread evolution of novel functions. However, high TE activity is not present in all tissues in the same way. Studies in maize showed that, although the host organism has mechanisms to prevent TE activity, TEs are in general at least low-expressed in all organs and under all conditions; only a few have high expression in certain organs. Interestingly, relatively high expression levels were detected in male reproductive organs, which can lead to transposable events in germ lines that might be therefore transferred to the next generation (Vicient, 2010). Most expressed TEs in germ lines belong to the Gypsy-like family, which could be a hint of the high abundance of this TE family in the genome of many plant species. Although epigenetic changes may explain the activity of several TE families in specific tissues, not all TE families are expressed in a similar manner, and other mechanisms might be involved in the specific activation of certain TE families. Additionally, several TEs are activated during stress responses. Possibly the most prominent example is the LTR retrotransposon *Tnt1* that could be activated through a cis-motif after wounding and viral, bacterial or fungal attacks (Grandbastien et al., 1997; Grandbastien, 1998).

All these results, including the findings presented in manuscript I and II, have strengthened the hypothesis that gene expression changes introduced by closely located
7.3 Transposable elements promote rapid expression divergence

TEs provided the raw material for the formation of promoter binding sites and significantly shaped the evolution of the nicotine biosynthetic pathway as well as herbivore-induced early defense signaling in *Nicotiana*. However, further experimental validation is needed to support this hypothesis by confirming that the newly introduced binding sites are indeed functional. To solve this question genome editing with the CRISPR/Cas system is a perfect way (Sander and Joung, 2014), e.g. by removing or modifying the proposed TF binding sites.

7.4 A combination of gene duplications and subsequent TE insertions directly contribute to the evolution of complex biosynthetic pathways

Specialized metabolites are remarkably diverse in plants. In total, more than 200,000 different chemical structures can be synthesized in plants (Hartmann, 2007), whereas each plant alone is already capable of synthesizing between 5,000 and 15,000 chemical structures (D’Auria and Gershenzon, 2005; Fernie, 2007). To illustrate the mechanism that contributed to the evolution of such highly diverse metabolites has been one of the central topics in plant biology. Previous studies, including the work I present in manuscript I and II, revealed that gene duplication followed by sub- or neofunctionalization can lead to novel enzymatic reactions and therefore novel metabolites and novel traits (Stehle et al., 2008; Wang et al., 2011; Hofberger et al., 2013). However, it remains difficult to detect direct evidence for the association to complex traits, such as multigenic pathways of secondary metabolites, though this association is theoretically intuitive. This remains difficult primarily due to the following reasons: (i) a detailed understanding of the ecological functions of traits is required; (ii) comprehensive information of the evolutionary history of participating genes and genomes is needed; and (iii) a broad spectrum of “-omics” data, such as transcriptomic and metabolomic data of different tissues, conditions and genotypes, is essential to link traits with their involved genes and metabolites. So far, only a few studies exist that show the direct contribution of duplications and TE insertions on the evolution of adaptive traits; however, this is mainly on traits correlated to single genes. One example is the evolution of threonine deaminase in the Solanaceae lineage. Higher organisms use this enzyme to catalyze the dehydration of threonine to α-ketobutyrate
and ammonia in the biosynthesis of isoleucine. But Solanaceae-specific duplications with subsequent expression and sequence divergence established a defense enzyme that acts in the gut of insects to degrade the essential amino acid threonine without being inhibited by isoleucine to disrupt their digestion in the herbivore (Gonzales-Vigil et al., 2011; Rausher and Huang, 2016). In manuscript I, we have further clarified this evolution and showed the explicit involvement of a TD event followed by a WGT-mediated duplication on this anti-digestive defense mechanism used in Solanaceae.

An additional example of the evolution of complex traits is C4 photosynthesis, which represents an advantage over the C3 carbon fixation common in most plants, and which represents an adaptation to warmer temperatures and to a lower CO₂ concentration (Ehleringer Helliher, 1997; Sage, 2004). Again, these studies suggest that the existence of redundant genes were a pre-requisite for C4 evolution (Monson, 2003; van den Bergh et al., 2014) and that subsequent transposon-mediated insertions of motifs might have played an important role in forming new cis-regulatory elements in the C4 photosynthesis-involved genes. However, additional experiments are needed to support these findings (Cao et al., 2016).

In summary, there is strong evidence that gene duplication and TE insertion are involved in the evolution of adaptive traits, although it is difficult to show their direct involvement. However, a single gene duplication or TE insertion does not create a new trait per se given their complexity, such as in the nicotine biosynthesis and in the early defense signaling as examined in manuscript I and II, respectively. Two concepts were proposed that might drive the evolution of adaptive traits in order to deal with herbivores. On the one hand, defense traits can evolve in a “gene-for-gene” co-evolution (Thompson and Burdon, 1992), indicating that the same genes experience adaptive substitutions in plant and herbivore. On the other hand, the “stepwise model” suggests that each partner gradually develops new traits step-by-step. The observed evolution of the nicotine biosynthesis described in manuscript I suggests the latter hypothesis might be involved: the stepwise model defined by the timing and the mode of duplications of involved genes reflects the diversification of alkaloid metabolism in the Solanaceae lineage. First, the genes ODC, PMT and MPO were duplicated and retained in Nicotiana, Petunia and Solanum, and represent the origin for the a common building blocks of most alkaloids in Solanaceae and Convolvulaceae, whereas in the Nicotiana
7.4 A combination of gene duplications and subsequent TE insertions directly contribute to the evolution of complex biosynthetic pathways

lineage this pathway was complemented at a later time point by a lineage-specific duplication of *AO, QPT* and *BBL*, with a potentially transposon-mediated gene expression divergence that completed this process to evolve a new alkaloid, nicotine, as a strong anti-herbivore defense compound. In summary, the evolution of nicotine biosynthesis might represent a perfect example for direct evolution of complex biosynthetic pathways in plant species.

7.5 *NaDH* to study adaptive traits in *Nicotiana*

Plants evolved a large number of mechanisms consisting of a broad genetic and metabolic repertoire to deal with biotic and abiotic stresses in their native habitats. Stress-induced defense responses often involve complex signaling transductions and transcriptional re-programing. To illustrate this process at the molecular level requires the integration of multi-“omics” data using different bioinformatics tools. In particular, a key aspect to understand how herbivore attack induces defensive responses requires the knowledge of how gene expression changes regulate the biosynthesis of defensive metabolites, as well as on how such complex regulation processes evolved. In particular, the combination of data from various high-throughput platforms such as genomics, phylogenetics, metabolomics and transcriptomics has a high potential for allowing the understanding of the “complete picture” and for overcoming limitations introduced by disparate data. For example, gene families tell us which genes might have similar functions, but usually genes need to interact in a specific context of other genes to fulfill their particular functions. On the other hand, new functions can arise through the modification of existing genes, e.g. by co-option (True and Carroll, 2002), which do not represent necessarily a similar functions as other genes from the same gene family.

As shown in several studies, including in manuscript 1, co-expression analysis is a powerful and widely-used tool for studying plant responses and for predicting functional associations between genes and their biological pathways (Bassel et al., 2011; Vanholme et al., 2012), as co-expression analysis can answer two important questions: (i) are the genes functional in a given metabolic pathway, and (ii) how are the genes controlled? Genes that are similar in expression among a set of experimental conditions, such as tissues, genotypes, biotic and abiotic stresses, are likely to be involved in the
same biochemical pathways (Aoki et al., 2007). Given that plant species synthesize between 5,000 and 15,000 chemical structures, and that the majority of these structures belong to structurally diverse secondary metabolites (D’Auria and Gershenzon, 2005; Fernie, 2007), increasing availability of non-targeted and large-scale metabolite profiling, such as large mass spectrometry libraries with several hundred metabolites, provides an important data source to understand the involved cellular processes and their functions, as many of them are specially produced under certain conditions that are still unknown. Current studies by Gulati and others (Hirai et al., 2004; Gulati et al., 2013a; Gulati et al., 2013b; Gulati et al., 2014; Li et al., 2016) showed that in particular co-expression between genes and metabolites provides detailed insight into the regulation and biosynthesis of secondary metabolites. To perform these analyses, however, complex mathematical and statistical knowledge is needed, including programming skills that are not achievable by every researcher, particularly not during fast hypothesis testing. At this point, online databases come into play and provide an easy and user-friendly access to this type of analysis. In manuscript III, I introduce a new database: *Nicotiana attenuata* Data Hub (NaDH). Although several databases for gene co-expression analyses exist, for example ALCOdb for microalgae (Aoki et al., 2016) and TFGD for tomato (Fei et al., 2011), they mainly focus on the prediction of gene functions based on known similarly expressed candidates; only a few exist which provide tools for understanding how this co-expression might have evolved. One example is FamNet (Ruprecht et al., 2016), which uses ortholog information. But manuscript I, II and III, including other studies (Gonzales-Vigil et al., 2011; Rausher and Huang, 2016) showed in particular the high potential of analyzing genome evolution, gene duplication and gene co-expression in parallel. This gap is now filled in NaDH by including phylogenetic and genomic datasets in a complex but intuitive visualization in order to combine gene co-expression and gene structure evolution. Aside from functional and expression similarities of genes, one key piece of information is gene age and gene duplication events in order to provide a fast understanding of the evolutionary history of co-expression networks. This can help to distinguish between different hypotheses of why duplicated copies are retained and conserved in the genome.
However, co-expression analysis should be used with caution, and has mainly three limitations that should be taken into consideration. First, transcripts, proteins and metabolites often have low correlation, mainly due to their different transport and regulation processes in cells. Therefore, co-expression is only meaningful among genes and metabolites that are expressed and synthesized in the same tissues where they accumulate, and which shows that co-expression does not always indicate co-function. Second, co-expression is the result of transcriptional and post-transcriptional regulation of mRNAs, and this relationship is often not known. Third, the observed co-expression can only detect the signals in the used data sets and is therefore highly dependent on them; larger data sets can provide more precise results. Forth, samples are often merged from different cell types although they come from the same tissues; in some occurrences, this can lead to false correlations.

Studies performing comparative expression analyses in plants revealed that several co-expression networks are conserved across species and kingdoms (Ficklin and Feltus, 2011; Gerstein et al., 2014; Zarrineh et al., 2014). Conserved expression modules were detected in many important biological processes, such as photosynthesis (Ficklin and Feltus, 2011), synthesis of secondary metabolites, and plant defenses (Humphry et al., 2010). This makes co-expression analysis a powerful framework for understanding the evolution of complex traits on a broad scale.

Taken together, by using comparative genomics of closely related species, my dissertation provides new evolutionary and mechanistic insights on how gene duplication and TE insertions are involved in the evolution of complex biosynthetic pathways and novel traits in general. Our results suggest that TEs play an essential role in remodeling and coordinating gene expression and deserve a higher focus in evolutionary biology.
8 References


Detzel A, Wink M (1993) Attraction, deterrence or intoxication of bees (Apis mellifera) by plant allelochemicals. Chemoecology 4: 8–18

189
8 References


Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A 106: 5737–42


References


Li D, Heiling S, Baldwin IT, Gaquerel E (2016) Illuminating a plant’s tissue-specific metabolic diversity using computational metabolomics and information theory. Proc Natl Acad Sci 201610218


8 References


8 References


Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. 934–946


9 Summary

Adaptation to local habitat is essential for the survival of all organisms, a process that is facilitated by adaptive traits that protect them from environmental stresses. To understand the evolution of adaptive traits has long been the central topic of evolutionary biology. However, most of adaptive traits evolved long time ago, elucidating their evolutionary history remains a big challenge, because it requires genomic information across closely related species, detailed knowledge on the genetic basis of the adaptive traits, and availability of sophisticated phylogenomic tools. Although gene duplication and transposable elements (TE) insertions are thought to be two important motors in the evolution of novel adaptive traits, details of the involved processes remains unclear. In this thesis, my work provides new insights on this fascinating topic.

First, I analyzed the evolution history of *Nicotiana* genomes by developing and applying bioinformatic tools of comparative genomics (Manuscript I). The results showed that genomes of *Nicotiana* species experienced a whole-genome triplication (WGT) event that is shared within the Solanaceae lineage and a repertoire of rapidly expanding TEs that bloated the *Nicotiana* genomes. To elucidate the evolution of nicotine biosynthetic machinery, phylogenomic analysis were performed and showed that the nicotine biosynthetic pathway gradually evolved from two duplicated ancient primary metabolic pathways followed by the rapid acquisition of root-specific gene expression. Analyses of the mechanisms of gene expression divergence revealed that the acquisitions of rapid coordinated expression were facilitated by different TE insertions that occurring after the Solanaceae whole-genome triplication event.

Second, the early defense signaling was analyzed in detail by using comparative transcriptomics and network analyses with six *Nicotiana* species (Manuscript II). This analysis revealed a key gene co-expression network that is co-activated with herbivore associated elicitors induced by jasmonate acid (JA) accumulations but independent from JA. Interestingly, the results showed again that numerous TE insertions in regulatory regions and gene retention after the WGT are involved and promoted the evolution of this key network. Their effects might be related with increased network complexity and robustness.
Finally, to provide free and easy access to the data and tools I used for my dissertation, I developed an interactive platform: the *Nicotiana attenuata Data Hub* (NaDH) (*Manuscript III*). Genomic, phylogenetic, transcriptomic and metabolomic data were used to form a comprehensive toolbox which not only allows users to explore and visualize the complex datasets and infer the putative functions of their gene of interest, it also facilitate reconstructions of other biosynthetic pathways in *Nicotiana*.

In conclusion, this dissertation sheds light on how the combination of gene duplications and TE insertions facilitate the evolution of multigene metabolic pathways in plants. The data and tools I established will further enhance *N. attenuata* as a model organism for studying plant-environment interactions and the evolution of adaptive traits in plants.
10 Zusammenfassung


Weiterhin untersuchte ich im Rahmen meiner Dissertation mittels vergleichende Transkriptom- und Netzwerkanalyse die rasche Signalaktivierung von


11 Eigenständigkeitserklärung


Jena, den 23.03.2017

__________________________
Thomas Brockmöller
12 Acknowledgment

First, I want to thank Prof. Ian T. Baldwin for believing in my expertise and giving me the opportunity to work in his department together with great scientists. His enthusiasm and his way of thinking provided a working atmosphere to not only understand but also to experience the impotence of science to solve fundamental questions and to extend my own limits. This will be very meaningful in my future life. Additionally, I want to thank him for his financial support to go to workshops for continuously learning new techniques and to improve my knowledge.

Furthermore, I want express my gratitude to my supervisor Dr. Shuqing Xu. His supervision and guidance during the last years shaped me fundamentally - as a person and specially as a scientist. All the skills, such as writing, presenting, scientific thinking, etc., that I have learned and improved with his help will be of big value for my whole life. I want to thank him for his encouraging words and discussions, when I get lost in analyses and data, and bring me back on track. I can’t thank him enough.

I’m very thankful to all current and former members of the Department of Chemical Ecology and thank them for their wonderful discussions and helpful hands. It was an honor working with them. Especially, I want to thank Zhihao for his computational expertise and to be a good friend.

In addition, I want to thank the IT department for their remarkably support and help in technical questions. A special thanks goes to Martin Niebergall for often dealing with my ‘specific requirements’.

Thanks to the International Max-Planck Research School (IMPRS) for its interesting lectures that increased my interdisciplinary thinking. A special thanks goes to coordinators Karin and Claudia which helped my a lot with bureaucratic matters.

I want to thank the Max-Planck Society for giving my the opportunity to do my PhD in such a good environment and for their financial support during my time as a PhD student.

A remarkable thanks goes to the ‘Mensa-groups’ for many funny but also serious discussions just about everything in the world. It was always a welcome refreshment besides the intensive work.
12 Acknowledgment

13 Curriculum Vitae

Personal data:
Name: Thomas Brockmöller

Education:

List of Publications:


List of Oral Presentations:

- Brockmöller, T. (2016). Comparative *Nicotiana* genomics shed lights light on the evolution of ecologically important traits. 15\textsuperscript{th} IMPRS Symposium, MPI for Chemical Ecology, Dornburg, Germany.

List of Poster Presentations:

- Brockmöller, T., Baldwin, I.T., Xu, S. (2014). Investigate the evolution of ecologically important traits in *Nicotiana* using a comparative genomics approach. 13\textsuperscript{th} IMPRS Symposium, MPI for Chemical Ecology, Dornburg, Germany.
- Gase, K., Brockmöller, T., Ling, Z., Pradhan, M., Sethi, A., Xu, S. (2016). The genome of *Nicotiana attenuata* and the genome-enabled approaches to ecology that it empowers; SAB Meeting, MPI for Chemical Ecology, Jena, Germany.

Jena, den 23.03.2017

Thomas Brockmöller
Comparative *Nicotiana* genomics shed lights on the evolution of adaptive traits in plants

by Thomas Brockmöller

**Theses:**

1. *N. attenuata* is a perfect model organisms to study the evolution of adaptive traits.

2. Gene duplication and transposable elements (TEs) are key drivers for genome evolution.

3. *Nicotiana* genomes are shaped by the rapid expansion of different TE families and the recent whole-genome triplication event (WGT) that is shared among Solanaceae.

4. Large-scale and small-scale gene duplications provide raw material for genomic innovations and contribute differently to the evolution of adaptive traits.

5. The nicotine biosynthesis in *Nicotiana* spp. evolved by a step-wise duplication of two primary metabolic pathways.

6. TE insertions and gene duplications highly contribute to gene expression divergence between gene pairs and therefore they promote neo- and subfunctionalization.

7. TE-induced transcription factor binding sites shaped and rewired the regulation network of the nicotine biosynthesis and the early defense signaling in *Nicotiana* spp.

8. A combined approach by using genomic, metabolomic and transcriptomic data including co-expression analysis has the potential for rapid discovery of gene and metabolite functions and its regulation.

9. The *Nicotiana attenuata* data hub (NaDH) is a resource to explore the function and evolution of genes and metabolites in *Nicotiana* species.