Regulation and evolution of alternative splicing 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

Zhihao Ling, M.S.

geboren am 10.08.1988 in China

Max-Planck-Institut für chemische Ökologie

Gutachter:

Prof. Ian T. Baldwin, Max Planck Institut für Chemische Ökologie, Jena

Prof. Günter Theissen, Friedrich Schiller Universität Jena

Prof. Dorothee Staiger, Universität Bielefeld

Beginn der Promotion: 26. September 2012

Dissertation eingereicht am: 25. November 2016

Tag der Verteidigung: 28. June 2017

Table of Contents

1. General Introduction

1.1 Alternative splicing is widespread in plants

Alternative splicing (AS) is a mechanism by which two or more transcripts are generated by removing different introns or using different splice sites from the same pre-mRNA. In general, AS can be broadly classified into four fundamental types (Figure 1): alternative donor sites (AltD), alternative acceptor sites (AltA), intron retention (IR) and exon skipping (ES) (Black 2003). A given splice variant may also contain combinations of different AS types at different introns, thus generating a more complex AS pattern. The identification of two different polypeptides from ribulosebisphosphate carboxylase/oxygenase (rubisco) activase in spinach and *Arabidopsis thaliana* provided the first demonstration of AS in plants (Werneke *et al.* 1989). Before next generation sequencing (NGS) technology was developed, AS was thought to be rare in plants, even though it was known to be common in humans (Brett *et al.* 2002). The estimations of AS frequency in *A. thaliana* using expressed sequence tags (ESTs) revealed that only 11% to 30% of multi-exon genes underwent AS and the discovery of AS depended highly on the sequencing depth of ESTs (Campbell *et al.* 2006, Iida *et al.* 2004). Recently, by sequencing cDNA derived from poly (A) enriched mRNA using NGS technology, several recent genome-wide AS studies in *A. thaliana* (Marquez *et al.* 2012), rice (Zhang *et al.* 2010), grape (Zenoni *et al.* 2010), soybean (Shen *et al.* 2014), moss (Wu *et al.* 2014) and wild tobacco (Ling *et al.* 2015) demonstrated that AS is much more prevalent in plants than previously thought, and that AS can be found in up to 66% of intron-containing genes.

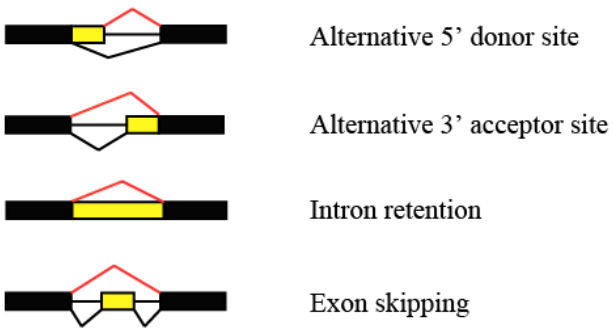


Figure 1. The four fundamental types of alternative splicing

The above mentioned genome-wide AS analyses all indicated that the predominant type of AS in land plants is IR, while ES has the lowest frequency. The pattern is opposite to that in metazoans, where ES is the most abundant AS type and IR is least prevalent (Kim *et al.* 2007, Pan *et al.* 2008). This indicates that the regulatory machinery of AS in plants might be different from metazoans. Previous studies demonstrated that introns from a human gene cannot be processed in plants but introns from a plant gene can be efficiently spliced in humans, providing further evidence to support this (Barta *et al.* 1986, Hartmuth and Barta 1986).

1.2 AS contributes to biological regulation processes in plants

One direct effect of AS is that the sequences of distinct splicing variants are different; if all these transcripts can be translated into proteins, proteome diversity can be expanded without corresponding expansion of the genome. In addition, the generated protein isoforms may also have different alterations, because AS can result in new transcripts that preserve the original open reading frame (if the difference caused by AS is in multiples of three nucleotides (nt)) or else cause a frameshift. The most commonly known non-frame shift AS is NAGNAG (N is any nucleotide) acceptor, which is a subset of AltA in which two splice sites (SS) are separated by three nt, thus resulting in two protein isoforms with only one amino acid difference (Iida *et al.* 2008, Schindler *et al.* 2008). In both plants and animals, NAGNAG events are enriched in genes encoding DNA-binding proteins and mainly influence polar amino acids that can change DNA-binding properties (Iida, et al. 2008, Schindler, et al. 2008, Vogan *et al.* 1996). When AS introduces reading frame shifts, it may generate either a protein isoform with novel functional domains or truncated protein isoforms, or else introduce premature termination codons (PTC) that result in transcript degradation. Protein isoforms with different domain arrangement may have different functions or differ in their subcellular localization and stability (Syed *et al.* 2012). The truncated proteins may lack certain functional domains or may have lost sites of post-translational modifications (Dinesh-Kumar and Baker 2000). For instance, several recent studies demonstrated that truncated proteins encoded by splicing variants can act as dominant-negative regulators by forming nonfunctional dimers that compete with functional dimers, which leads to physiological responses (Liu *et al.* 2013, Seo *et al.* 2012, Staudt and Wenkel 2011). However, functional studies on the proteins generated by AS are rare in plants, and some authors also raise

the question of whether AS really played an important role in expanding proteome diversity in plants (Severing *et al.* 2009).

As mentioned, another important effect of AS is that it can regulate transcript abundance by introducing PTC to transcripts, which may become subject to degradation by nonsense-mediated decay (NMD), a cytoplasmic RNA degradation machinery (Chang *et al.* 2007, Kalyna *et al.* 2012, Kervestin and Jacobson 2012, Schoenberg and Maquat 2012). NMD initiates rapid transcript decay when the termination of translation is perturbed, although, the detailed NMD pathway in plants is not yet well characterized (Belostotsky and Sieburth 2009, Chang, et al. 2007, Schoenberg and Maquat 2012). The orthologues of the core proteins of the NMD pathway, UPF1-3 and SMG-7 can be found in most of plants (but not SMG-1, SMG-5 and SMG-6), suggesting the NMD machinery is largely conserved (Arciga-Reyes *et al.* 2006, Hori and Watanabe 2005, Kerenyi *et al.* 2008, Riehs *et al.* 2008). NMD-sensitive transcripts in plants have common features including: (i) a PTC occurs more than 50-55 nt upstream of a splice junction or is located in the middle of the coding sequence (Kerenyi, et al. 2008, Wu *et al.* 2007b); (ii) have a long 3' UTR (> 350 nt) or have introns in the 3' UTR (Kalyna, et al. 2012); (iii) contain upstream open reading frames (uORFs) which overlapped the start codon (AUG) of the main ORF (Kalyna, et al. 2012, Nyiko *et al.* 2009). AS can introduce all of these features to the transcripts, thereby affecting their stability. Many splicing factors (SFs) in plants are found to produce PTC+ transcripts that are involved in the autoregulation of their own or other SFs' expression through a negative feedback loop (Kalyna *et al.* 2006, Stauffer *et al.* 2010, Wachter *et al.* 2012). However, it has also been found that some AS transcripts with NMD features were not sensitive to NMD, especially for transcripts with IR, suggesting there is a specific strategy to avoid NMD for some PTC+ transcripts in plants (Kalyna, et al. 2012, Leviatan *et al.* 2013). Two recent studies revealed that some PTC+ intron-containing transcripts can have much longer half-lives in the nucleus, thus allowing them to escape the NMD machinery (Boutz *et al.* 2015, Gohring *et al.* 2014).

Furthermore, AS is also involved in microRNA (miRNA) associated regulation of mRNA stability. miRNA is a class of endogenous RNAs with a length of ~23 nt length that play important regulatory roles in gene expression (GE) at post transcriptional level by pairing to partial complementary miRNA binding sites (MBS) in their target mRNAs (Bartel 2009). There are two ways that AS can modulate miRNA-mediated regulation of GE. Firstly, AS can generate

transcript isoforms that either lack or contain related MBS, thereby excluding or receiving the miRNA regulation. For example, the splicing variants of transacting small interfering RNA (*TAS*) genes and squamosa-promoter binding protein like-4 (*SPL4*) gene in *A. thaliana* that do not contain the MBS of their related miRNA enables the escaping of miRNA regulation that would silence the MBS contained splicing variants (Wu and Poethig 2006, Yang *et al.* 2012b). In addition, it was found that over 12% of identified MBS in *A. thaliana* are affected by AS, suggesting the association between AS and miRNA-mediated gene regulations are not rare (Yang, et al. 2012b). Secondly, AS in pre-mRNA regions, where generate miRNA can also modulate the abundance of the same miRNA, which in turn modulates the stability of their target mRNAs. For instance, *MIR400*, a miRNA located in the first intron of a protein coding gene in *A. thaliana*, shows consistent expression pattern with its host gene. In response to heat stress, the abundance of splicing variant retaining the intron containing *MIR400* was up-regulated, and thus the expression of mature miR400 was suppressed. This splicing change appears to be important for the tolerance of heat stress, because under heat stress, the overexpression of miR400 resulted in decreased seedling growth and germination under heat stress (Yan *et al.* 2012). Furthermore, miRNA can locate both in intronic and exonic regions of their host genes (Yang *et al.* 2012a), suggesting the correlation between miRNA and AS may be underestimated. However, the overall roles of AS in miRNA-mediated regulation of GE remain unclear.

1.3 Abiotic stresses induced AS changes in plants

Being sessile, plants must respond to environmental stresses in order to adapt and survive under diverse conditions. To deal with abiotic stresses and maximize their fitness, plants regulate their transcriptome rapidly at multiple levels, including modifying the expression and splicing of stress-related genes (Ali and Reddy 2008a, Staiger and Brown 2013, Yamaguchi-Shinozaki and Shinozaki 2006). Among various abiotic stresses, high and low temperature stresses are probably the most commonly known and best-studied stresses in plants Both elevated and decreased temperature can induce genome-wide modifications of AS in plants, although the range of the changes is relatively small (Chang *et al.* 2014, Leviatan, et al. 2013, Streitner *et al.* 2013). Case studies on key regulatory genes further confirmed significant changes of AS in response to temperature changes, suggesting the AS in response to temperature can serve as ‘molecular thermometers’ (Capovilla *et al.* 2015). For example, AS is regulated differentially under elevated

temperatures, since a subsequent heat shock (HS) treatment has a moderate effect on AS compared with the initial HS, indicating AS also plays a role in thermotolerance responses (Chang, et al. 2014). This study further identified an exonic GAC-repeat motif, which may act as a putative regulatory *cis*-element in heat-mediated IR regulation (Chang, et al. 2014). The splicing responses of the heat shock transcription factor A2 (*HsfA2*) in *A. thaliana* is another example. *HsfA2* can generate two alternatively spliced transcripts in response to various temperatures: while splice variant *HsfA2-II* is generated at moderate heat (37 °C) and decreases along with elevated temperature, *HsfA2-III* only appears under severe HS (≥42°C) (Liu, et al. 2013, Sugio *et al.* 2009). Interestingly, *HsfA2-II* contains a PTC and is degraded by NMD but *HsfA2-III* can encode a truncated protein, S-HsfA2, which can bind to the heat shock elements (HSE) in the *HsfA2* promoter, thus indicating an auto-regulatory loop to activate its own transcription (Liu, et al. 2013). Cold stress also induces changes of the AS of several serine/arginine-rich (SR) protein encoded genes and core circadian clock genes, which are involve in the regulation of splicing, physiological and developmental processes in plants (Capovilla, et al. 2015, James *et al.* 2012, Tanabe *et al.* 2007).

In addition to temperature fluctuations, plants also face many other abiotic stresses such as salt, light irradiation, drought and nutrient deficiency. Several recent studies have revealed that these stresses can also globally induce significant AS changes in plants (Ding *et al.* 2014, Li *et al.* 2013, Thatcher *et al.* 2016, Wu, et al. 2014). In general, several common features have been observed in plants under various stress conditions (including temperature fluctuations): (i) the total number of AS events are increased under salt stress and nutrient deficient conditions compared with normal conditions (Ding, et al. 2014, Li, et al. 2013, Zhang *et al.* 2016); (ii) AS and GE are separately regulated in response to stress (Ding, et al. 2014, Li, et al. 2013); (iii) similar regulatory *cis*-elements might be involved in regulating AS in response to different stress conditions, since a purine rich (GAA-repeat) exonic motif was identified in moss that potentially functions in regulating light-mediated IR (Wu, et al. 2014). This motif is consistent with above mentioned exonic motif, which may be involved in the regulation of heat-induced IR (Chang, et al. 2014); (iv) the AS patterns of splicing factors (mainly SR proteins) and core circadian clock genes have all been shown to have significant changes in response to various stresses (Ding, et al. 2014, Filichkin *et al.* 2010, Palusa *et al.* 2007, Syed *et al.* 2015, Tanabe, et al. 2007, Wu, et al.

2014). These studies suggest that different stress-induced AS may share a common regulatory pathway.

The mechanism of how stress signals reprogram the splicing of the transcriptome is largely unknown in plants. Based on previous studies, stress signals may regulate AS through several potential mechanisms (Figure 2). Firstly, stresses induce changes of expression and AS in the splicing regulators (e.g. SRs) that directly lead to global changes in AS (Filichkin, et al. 2010, Gulledge *et al.* 2012, Lee *et al.* 2006, Palusa, et al. 2007). Secondly, stresses that induce phosphorylation or transcriptional changes of protein kinases that phosphorylate SFs can also result in global AS regulation (Ali and Reddy 2008b, Monks *et al.* 2001). Several plant splicing factors (SFs) are known to be phosphoproteins, and the phosphorylation status of SR proteins can affect their mobility, subcellular localization and may even affect their functions (Ali *et al.* 2003, de la Fuente van Bentem *et al.* 2008, Tillemans *et al.* 2005, Tillemans *et al.* 2006, Zhang and Mount 2009). Thirdly, stresses induce RNA structure changes. It has been revealed that in yeast, RNA folds can control the alternative 3' splice site (SS) availability, which affect AS in response to heat shock (Meyer *et al.* 2011). This mechanism is also not well studied in plants.

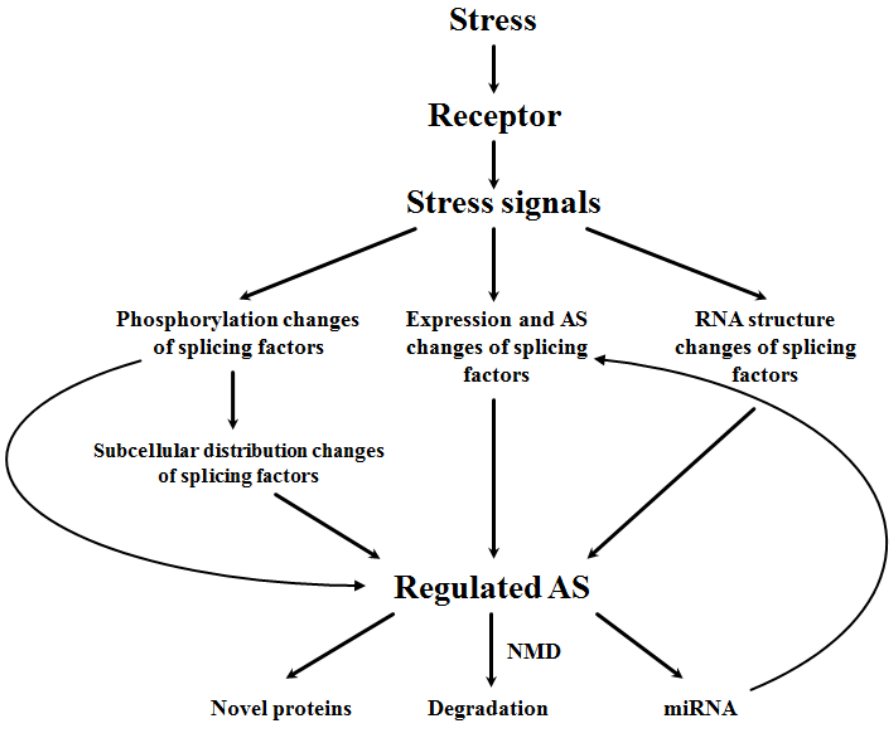


Figure 2. The potential mechanisms of how stress signals regulate AS in plants.

1.4 Biotic stresses induced AS changes in plants

Besides abiotic stress, plants are also confronted with many biotic stresses, of which pathogens and herbivores are the two main types. In response to pathogen infection, the AS regulations of resistance (R) genes plays an important role in plant defense by generating R proteins that provide plant disease resistance against pathogens, (Dinesh-Kumar and Baker 2000, Zhang and Gassmann 2003, Zhang and Gassmann 2007). In addition, plant resistance to pathogens is affected by the relative ratios of multiple transcripts, especially infection-induced transcripts (Dinesh-Kumar and Baker 2000, Zhang and Gassmann 2007). For example, the AS of some R genes generates transcripts with PTC that could be subject to NMD or translated into truncated proteins. However, the presence of both transcripts encoding full-length and truncated ORFs is required for full resistance (Dinesh-Kumar and Baker 2000, Zhang and Gassmann 2003, Zhang and Gassmann 2007). Interestingly, it has been found that in *Pst* DC3000-infected plants, the expression of genes involved in the NMD pathway, such as *UPF1* and *UPF3*, were down-regulated, which enabled the PTC+ transcripts of R gene to not be degraded by NMD and to produce truncated proteins that are involved in induced cell death (Jeong *et al.* 2011, Michael Weaver *et al.* 2006, Zhang and Gassmann 2007). At the genome-wide level, a recent study showed that pathogen infection induced large numbers of novel AS in *A. thaliana*, including AS of genes that are likely to play important roles in defending against invasive pathogens (Howard *et al.* 2013). However, in comparison to abiotic stress, biotic stress induced AS responses in plants are largely unexplored, particularly the regulation of AS in response to herbivores. In **manuscript I**, I investigated *Manduca sexta* (a specialist herbivore of Solanaceae) induced genome-wide AS responses in wild tobacco (*Nicotiana attenuata*), in both local tissue (leaves) and systemic tissue (roots).

Nicotiana attenuata, is native to the Great Basin Desert in the United States and occurs primarily in large ephemeral populations in post-fire habitats, and is therefore subject to highly dynamic biotic environmental pressures to re-establish their populations (Bahulikar *et al.* 2004, Baldwin *et al.* 1994). The adaptability of *N. attenuata* to environmental changes is mediated by a high degree of phenotypic plasticity, as illustrated by sophisticated direct and indirect defense mechanisms induced by herbivores (Kessler *et al.* 2004, Wu and Baldwin 2010, Wu *et al.* 2007a). In addition, a high-quality genome of *N. attenuata* has been sequenced and assembled

(**manuscript II**). All these features make *N. attenuata* an excellent model system for studying herbivore induced AS in plants.

1.5 The splicing code and determinants of AS in plants is largely unknown

The spliceosome is a highly dynamic RNA-protein complex consisting of five small nuclear ribonucleoprotein particles (snRNPs) and over 300 proteins in humans (Jurica and Moore 2003, Rappsilber *et al.* 2002), but the spliceosome in plants has not been isolated so far (Reddy *et al.* 2013). Assembly of the spliceosome to remove introns and ligate exons is directed by sequence features of the pre-mRNA; however, how the exonic and intronic regions in plants are recognized is poorly understood. In metazoans, it is well known that four crucial signals are required for accurate splicing including (i) 5' splice sites (SS), which contain a GU dinucleotide at the intron start that is recognized by U1 snRNP, (ii) 3' SS, which include an AG at the extreme 3' end that is recognized by U2 auxillary factor (U2AF35), (iii) a polypyrimidine tract (PPT) located at the 3' end of the intron that is recognized by U2AF65, and (iv) a branch site (BS) sequence located ~17-40 nt upstream of the 3' SS that is recognized by U2 snRNP (Black 2003). Except BS, similar sequence features were found in plants but with some minor specific nt frequency difference at specific positions (Reddy 2007). In addition, a UA-rich tract in introns was also found to be important for efficient splicing in plants (Baek *et al.* 2008, Lewandowska *et al.* 2004, Simpson *et al.* 2004). For splice site recognition, two models named 'exon definition' and 'intron definition' have been proposed (Berget 1995). In general, 'exon definition' is favored in organisms with relatively small exons flanked by large introns, while 'intron definition' is favored in organisms with relatively small introns (Sterner *et al.* 1996, Talerico and Berget 1994). In plants, it is proposed that both models are used to recognize SS with predominant preference in 'intron definition' (Brown 1996, Lorkovic *et al.* 2000). Furthermore, Renddy (2007) hypothesized that different composition of introns between rice and *A. thaliana*, which have different intron sizes and GC contents, may have different splicing machineries.

The regulation of AS depends largely on *cis* signals close to the splicing site (SS) and *trans*-acting splicing factors (SFs) that can recognize the signals. In plants, there are over 300 RNA-binding proteins that have one or more RNA-binding domains, such as RNA recognition motif (RRM) and the K homology domain, which may function as splicing regulators (Silverman

et al. 2013). Among them, SR proteins are the most intensively studied and are known to regulat AS (Gao *et al.* 2004, Lopato *et al.* 1999a, Lopato *et al.* 1999b). Interestingly, several studies found that SR proteins can auto-regulate their own AS or that of other SR genes (Isshiki *et al.* 2006, Kalyna, et al. 2006, Thomas *et al.* 2012). Among eukaryotes, plants have the highest number of SR proteins, almost double the number of non-photosynthetic organisms (Richardson *et al.* 2011). However, the number of SR proteins varies among different plant species (24 in rice and 19 in *Arabidopsis*) (Iida and Go 2006, Isshiki, et al. 2006). The detailed functions of each SR protein in plants and whether different number of SR proteins contributes to species-specific AS profiles is largely unclear.

In animals, many splicing regulatory elements (SREs) and RNA binding proteins (RBPs) have been identified using different strategies, and the interactions between these SREs in the pre-mRNA and RBPs can either promote or suppress the use of a splice site (Barash *et al.* 2010, Chen and Manley 2009, Licatalosi *et al.* 2008). These SREs can be generally classified as exonic splicing enhancers (ESEs)/silencers (ESSs) and intronic splicing enhancers (ISEs)/silencers (ISSs), which are short (3-11 nt) conserved sequences that often occur in clusters (Chen and Manley 2009). Some of them can be used to predict the outcomes of specific AS type, which indicates these *cis*-elements are important factors in regulating AS (Barash, et al. 2010). While plants have, more than 200 RBPs, most of the identified being plant-specific, their binding sequence motifs are largely unknown (Lorkovic 2009). In addition, although over 80 SREs in plants have been identified in *A. thaliana*, only a few of them have been confirmed by experiments (Pertea *et al.* 2007, Schoning *et al.* 2008, Thomas, et al. 2012, Yoshimura *et al.* 2002).

Furthermore, some other factors including secondary and tertiary RNA structures, chromatin remodeling, insertion of transposable elements (TEs) and gene duplication (GD) have been found to affect AS regulations in animal systems but are largely unexplored in plants (Donahue *et al.* 2006, Kolasinska-Zwierz *et al.* 2009, Lambert *et al.* 2015, Liu *et al.* 1995, Schwartz *et al.* 2009, Sorek *et al.* 2002, Su *et al.* 2006, Warf and Berglund 2010). Furthermore, studies in mammals showed that the emergence of AS is originated from constitutive splicing through the fixation of SREs and the creation of alternative competing SS (Koren *et al.* 2007, Lev-Maor *et al.* 2007). There are distinct features between alternatively spliced exons/introns and constitutively spliced exons/introns, which can be used to accurately predict specific AS type

(Braunschweig *et al.* 2014, Koren, et al. 2007). However, these distinct features between AS and constitutive splicing remain unstudied in plants. In **manuscript III**, using a machine learning approach, I identified key features that contribute to the determination of AltA and AltD, which can be used to predict these two AS types in plants.

1.6 The rapid evolution of AS

Although AS is found in all eukaryotes, it is much more prevalent in highly evolved multicellular eukaryotes compares to ancestral unicellular eukaryotes. For example, while up to 60% of intron-containing genes underwent AS in plants, AS is rare or may be inexistent in budding yeast (Barrass and Beggs 2003, Eckardt 2002, Ling, et al. 2015, Reddy 2007). Based on the comparison of the 5' SS between unicellular and multicellular eukaryotes, which reveals a higher plasticity in the 5'SS of multicellular organisms than unicellular organisms, two AS evolution models were proposed: (i) sequence based model, in which the production of weak SS through the mutations in a constitutive SS provides the chance for the splicing machinery to skip the SS, thereby generating AS, and (ii) *trans*-factor based model, which invokes the evolution of SFs such as SR proteins and RBPs that releases the selective pressure on SS, resulting in mutations that weaken the SS (Ast 2004).

Soon after the discovery of AS in eukaryotic genes, AS was speculated to accelerate evolution (Gilbert 1978). This hypothesis is supported by two recent extensive studies in phylogenetically divergent vertebrate lineages, which reveal that AS may contribute to the speciation (Barbosa-Morais *et al.* 2012, Merkin *et al.* 2012). Both studies found that while tissue-specific gene expression (GE) patterns are largely conserved, AS profiles are frequently lineage-specific and are primarily *cis*-directed (Barbosa-Morais, et al. 2012, Merkin, et al. 2012). In plants, such in-depth comprehensive analyses to study AS evolution have not been performed. Most of studies have been focused only on specific genes or gene families which contain conserved AS between several species (Iida and Go 2006, Li *et al.* 2006). In plants, some genome-wide studies using expressed sequence tags (ESTs) in plants only identified a small number of conserved AS (Darracq and Adams 2013, Severing, et al. 2009, Wang and Brendel 2006, Wang *et al.* 2008), whereas other studies identified a much higher number of conserved AS between closely related species (Chamala *et al.* 2015, Satyawan *et al.* 2016). However, none

of these studies systematically investigated the evolution of AS and underlying mechanisms. In **manuscript III**, I analyzed transcriptomes of three tissues in six plant species to investigate the pattern of AS evolution in plants. Using a machine learning approach, I further identified the key factors that contributed to the determinations of AltA and AltD and rapid changes of AS between closely related species in plants.

1.7 Thesis outline

While stresses are known to induce AS changes in plants, how insect herbivory elicits AS responses in plants is unknown. Furthermore, the close association between AS and environmental stresses predicts that AS evolved rapidly in plants. However, evidence for this rapid evolution is limited and mechanisms remain unknown. The main objectives of my thesis are I) to investigate genome-wide herbivore-induced AS in plants, using wild tobacco as model system. Here, I aimed to address four main questions: a) to what extent can insect herbivores induce AS responses in wild tobacco? b) Are gene expression and alternative splicing co-regulated in response to herbivory? c) Do local and systemic tissues show different AS responses under herbivore attack? d) Is JA signaling involved in herbivory induced AS responses? II) To systematically investigate the patterns of AS evolution among different plant species and the mechanisms that contribute to changes of AS between closely related species. Here, I aim to address four main questions: a) Are AS profiles in plants species-specific? b) Does AS evolve faster than gene expression? c) What are the determinants of AS in plants? d) Which factors contributed to the evolution of AS?

References

Ali, G.S., Golovkin, M. and Reddy, A.S. (2003) Nuclear localization and in vivo dynamics of a plant-specific serine/arginine-rich protein. *Plant J*, **36**, 883-893.

Ali, G.S. and Reddy, A.S. (2008a) Regulation of alternative splicing of pre-mRNAs by stresses. *Curr Top Microbiol Immunol*, **326**, 257-275.

Ali, G.S. and Reddy, A.S. (2008b) Spatiotemporal organization of pre-mRNA splicing proteins in plants. *Curr Top Microbiol Immunol*, **326**, 103-118.

Arciga-Reyes, L., Wootton, L., Kieffer, M. and Davies, B. (2006) UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in Arabidopsis. *Plant J*, **47**, 480-489.

Ast, G. (2004) How did alternative splicing evolve? *Nat Rev Genet*, **5**, 773-782.

Baek, J.M., Han, P., Iandolino, A. and Cook, D.R. (2008) Characterization and comparison of intron structure and alternative splicing between *Medicago truncatula*, *Populus trichocarpa*, Arabidopsis and rice. *Plant Mol Biol*, **67**, 499-510.

Bahulikar, R.A., Stanculescu, D., Preston, C.A. and Baldwin, I.T. (2004) ISSR and AFLP analysis of the temporal and spatial population structure of the post-fire annual, *Nicotiana attenuata*, in SW Utah. *BMC Ecol*, **4**, 12.

Baldwin, I.T., Staszak-Kozinski, L. and Davidson, R. (1994) Up in smoke: I. Smoke-derived germination cues for postfire annual,*Nicotiana attenuata* torr. Ex. Watson. *J Chem Ecol*, **20**, 2345-2371.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., et al. (2010) Deciphering the splicing code. *Nature*, **465**, 53-59.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., et al. (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587-1593.

Barrass, J.D. and Beggs, J.D. (2003) Splicing goes global. *Trends Genet*, **19**, 295-298.

Barta, A., Sommergruber, K., Thompson, D., Hartmuth, K., Matzke, M.A. and Matzke, A.J. (1986) The expression of a nopaline synthase - human growth hormone chimaeric gene in transformed tobacco and sunflower callus tissue. *Plant Mol Biol*, **6**, 347-357.

Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215-233.

Belostotsky, D.A. and Sieburth, L.E. (2009) Kill the messenger: mRNA decay and plant development. *Curr Opin Plant Biol*, **12**, 96-102.

Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J Biol Chem*, **270**, 2411-2414.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**, 291-336.

Boutz, P.L., Bhutkar, A. and Sharp, P.A. (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev*, **29**, 63-80.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., et al. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res*, **24**, 1774-1786.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J. and Bork, P. (2002) Alternative splicing and genome complexity. *Nat Genet*, **30**, 29-30.

Brown, J.W. (1996) Arabidopsis intron mutations and pre-mRNA splicing. *Plant J*, **10**, 771-780.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. and Buell, C.R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics*, **7**, 327.

Capovilla, G., Pajoro, A., Immink, R.G. and Schmid, M. (2015) Role of alternative pre-mRNA splicing in temperature signaling. *Curr Opin Plant Biol*, **27**, 97-103.

Chamala, S., Feng, G., Chavarro, C. and Barbazuk, W.B. (2015) Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. *Front Bioeng Biotechnol*, **3**, 33.

Chang, C.Y., Lin, W.D. and Tu, S.L. (2014) Genome-Wide Analysis of Heat-Sensitive Alternative Splicing in *Physcomitrella patens*. *Plant Physiol*, **165**, 826-840.

Chang, Y.F., Imam, J.S. and Wilkinson, M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*, **76**, 51-74.

Chen, M. and Manley, J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*, **10**, 741-754.

Darracq, A. and Adams, K.L. (2013) Features of evolutionarily conserved alternative splicing events between Brassica and Arabidopsis. *New Phytol*, **199**, 252-263.

de la Fuente van Bentem, S., Anrather, D., Dohnal, I., Roitinger, E., Csaszar, E., Joore, J., et al. (2008) Site-specific phosphorylation profiling of Arabidopsis proteins by mass spectrometry and peptide chip analysis. *J Proteome Res*, **7**, 2458-2470.

Dinesh-Kumar, S.P. and Baker, B.J. (2000) Alternatively spliced N resistance gene transcripts: their possible role in tobacco mosaic virus resistance. *Proc Natl Acad Sci U S A*, **97**, 1908-1913.

Ding, F., Cui, P., Wang, Z., Zhang, S., Ali, S. and Xiong, L. (2014) Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. *BMC Genomics*, **15**, 431.

Donahue, C.P., Muratore, C., Wu, J.Y., Kosik, K.S. and Wolfe, M.S. (2006) Stabilization of the tau exon 10 stem loop alters pre-mRNA splicing. *J Biol Chem*, **281**, 23302-23306.

Eckardt, N.A. (2002) Alternative splicing and the control of flowering time. *Plant Cell*, **14**, 743-747.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., et al. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res*, **20**, 45-58.

Gao, H., Gordon-Kamm, W.J. and Lyznik, L.A. (2004) ASF/SF2-like maize pre-mRNA splicing factors affect splice site utilization and their transcripts are alternatively spliced. *Gene*, **339**, 25-37.

Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.

Gohring, J., Jacak, J. and Barta, A. (2014) Imaging of endogenous messenger RNA splice variants in living cells reveals nuclear retention of transcripts inaccessible to nonsense-mediated decay in Arabidopsis. *Plant Cell*, **26**, 754-764.

Gulledge, A.A., Roberts, A.D., Vora, H., Patel, K. and Loraine, A.E. (2012) Mining *Arabidopsis thaliana* RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. *Am J Bot*, **99**, 219-231.

Hartmuth, K. and Barta, A. (1986) In vitro processing of a plant pre-mRNA in a HeLa cell nuclear extract. *Nucleic Acids Res*, **14**, 7513-7528.

Hori, K. and Watanabe, Y. (2005) UPF3 suppresses aberrant spliced mRNA in Arabidopsis. *Plant J*, **43**, 530-540.

Howard, B.E., Hu, Q., Babaoglu, A.C., Chandra, M., Borghi, M., Tan, X., et al. (2013) High-throughput RNA sequencing of pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants. *PLoS One*, **8**, e74183.

Iida, K. and Go, M. (2006) Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. *Mol Biol Evol*, **23**, 1085-1094.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., et al. (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res*, **32**, 5096-5103.

Iida, K., Shionyu, M. and Suso, Y. (2008) Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals. *Mol Biol Evol*, **25**, 709-718.

Isshiki, M., Tsumoto, A. and Shimamoto, K. (2006) The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. *Plant Cell*, **18**, 146-158.

James, A.B., Syed, N.H., Bordage, S., Marshall, J., Nimmo, G.A., Jenkins, G.I., et al. (2012) Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *Plant Cell*, **24**, 961-981.

Jeong, H.J., Kim, Y.J., Kim, S.H., Kim, Y.H., Lee, I.J., Kim, Y.K., et al. (2011) Nonsense-mediated mRNA decay factors, UPF1 and UPF3, contribute to plant defense. *Plant Cell Physiol*, **52**, 2147-2156.

Jurica, M.S. and Moore, M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell*, **12**, 5-14.

Kalyna, M., Lopato, S., Voronin, V. and Barta, A. (2006) Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. *Nucleic Acids Res*, **34**, 4395-4405.

Kalyna, M., Simpson, C.G., Syed, N.H., Lewandowska, D., Marquez, Y., Kusenda, B., et al. (2012) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res*, **40**, 2454-2469.

Kerenyi, Z., Merai, Z., Hiripi, L., Benkovics, A., Gyula, P., Lacomme, C., et al. (2008) Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. *EMBO J*, **27**, 1585-1595.

Kervestin, S. and Jacobson, A. (2012) NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol*, **13**, 700-712.

Kessler, A., Halitschke, R. and Baldwin, I.T. (2004) Silencing the jasmonate cascade: induced plant defenses and insect populations. *Science*, **305**, 665-668.

Kim, E., Magen, A. and Ast, G. (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, **35**, 125-131.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S. and Ahringer, J. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet*, **41**, 376-381.

Koren, E., Lev-Maor, G. and Ast, G. (2007) The emergence of alternative 3' and 5' splice site exons from constitutive exons. *PLoS Comput Biol*, **3**, e95.

Lambert, M.J., Cochran, W.O., Wilde, B.M., Olsen, K.G. and Cooper, C.D. (2015) Evidence for widespread subfunctionalization of splice forms in vertebrate genomes. *Genome Res*, **25**, 624-632.

Lee, B.H., Kapoor, A., Zhu, J. and Zhu, J.K. (2006) STABILIZED1, a stress-upregulated nuclear protein, is required for pre-mRNA splicing, mRNA turnover, and stress tolerance in Arabidopsis. *Plant Cell*, **18**, 1736-1749.

Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., et al. (2007) The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet*, **3**, e203.

Leviatan, N., Alkan, N., Leshkowitz, D. and Fluhr, R. (2013) Genome-wide survey of cold stress regulated alternative splicing in *Arabidopsis thaliana* with tiling microarray. *PLoS One*, **8**, e66511.

Lewandowska, D., Simpson, C.G., Clark, G.P., Jennings, N.S., Barciszewska-Pacak, M., Lin, C.F., et al. (2004) Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell*, **16**, 1340-1352.

Li, J., Li, X., Guo, L., Lu, F., Feng, X., He, K., et al. (2006) A subgroup of MYB transcription factor genes undergoes highly conserved alternative splicing in Arabidopsis and rice. *J Exp Bot*, **57**, 1263-1273.

Li, W., Lin, W.D., Ray, P., Lan, P. and Schmidt, W. (2013) Genome-wide detection of condition-sensitive alternative splicing in Arabidopsis roots. *Plant Physiol*, **162**, 1750-1763.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., et al. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464-469.

Ling, Z., Zhou, W., Baldwin, I.T. and Xu, S. (2015) Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*. *Plant J*, **84**, 228-243.

Liu, H.X., Goodall, G.J., Kole, R. and Filipowicz, W. (1995) Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in *Nicotiana plumbaginifolia*. *EMBO J*, **14**, 377-388.

Liu, J., Sun, N., Liu, M., Liu, J., Du, B., Wang, X., et al. (2013) An autoregulatory loop controlling Arabidopsis HsfA2 expression: role of heat shock-induced alternative splicing. *Plant Physiol*, **162**, 512-521.

Lopato, S., Gattoni, R., Fabini, G., Stevenin, J. and Barta, A. (1999a) A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. *Plant Mol Biol*, **39**, 761-773.

Lopato, S., Kalyna, M., Dorner, S., Kobayashi, R., Krainer, A.R. and Barta, A. (1999b) atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes. *Genes Dev*, **13**, 987-1001.

Lorkovic, Z.J. (2009) Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends Plant Sci*, **14**, 229-236.

Lorkovic, Z.J., Wieczorek Kirk, D.A., Lambermon, M.H. and Filipowicz, W. (2000) Pre-mRNA splicing in higher plants. *Trends Plant Sci*, **5**, 160-167.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res*, **22**, 1184-1195.

Merkin, J., Russell, C., Chen, P. and Burge, C.B. (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593-1599.

Meyer, M., Plass, M., Perez-Valle, J., Eyras, E. and Vilardell, J. (2011) Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol Cell*, **43**, 1033-1039.

Michael Weaver, L., Swiderski, M.R., Li, Y. and Jones, J.D. (2006) The *Arabidopsis thaliana* TIR-NB-LRR R-protein, RPP1A; protein localization and constitutive activation of defence by truncated alleles in tobacco and Arabidopsis. *Plant J*, **47**, 829-840.

Monks, D.E., Aghoram, K., Courtney, P.D., DeWald, D.B. and Dewey, R.E. (2001) Hyperosmotic stress induces the rapid phosphorylation of a soybean phosphatidylinositol transfer protein homolog through activation of the protein kinases SPK1 and SPK2. *Plant Cell*, **13**, 1205-1219.

Nyiko, T., Sonkoly, B., Merai, Z., Benkovics, A.H. and Silhavy, D. (2009) Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol Biol*, **71**, 367-378.

Palusa, S.G., Ali, G.S. and Reddy, A.S. (2007) Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J*, **49**, 1091-1107.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **40**, 1413-1415.

Pertea, M., Mount, S.M. and Salzberg, S.L. (2007) A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. *BMC Bioinformatics*, **8**, 159.

Rappsilber, J., Ryder, U., Lamond, A.I. and Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome. *Genome Res*, **12**, 1231-1245.

Reddy, A.S. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol*, **58**, 267-294.

Reddy, A.S., Marquez, Y., Kalyna, M. and Barta, A. (2013) Complexity of the alternative splicing landscape in plants. *Plant Cell*, **25**, 3657-3683.

Richardson, D.N., Rogers, M.F., Labadorf, A., Ben-Hur, A., Guo, H., Paterson, A.H., et al. (2011) Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into sub-family classification and extent of alternative splicing. *PLoS One*, **6**, e24542.

Riehs, N., Akimcheva, S., Puizina, J., Bulankova, P., Idol, R.A., Siroky, J., et al. (2008) Arabidopsis SMG7 protein is required for exit from meiosis. *J Cell Sci*, **121**, 2208-2216.

Satyawan, D., Kim, M.Y. and Lee, S.H. (2016) Stochastic alternative splicing is prevalent in mungbean (*Vigna radiata*). *Plant Biotechnol J*.

Schindler, S., Szafranski, K., Hiller, M., Ali, G.S., Palusa, S.G., Backofen, R., et al. (2008) Alternative splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes. *BMC Genomics*, **9**, 159.

Schoenberg, D.R. and Maquat, L.E. (2012) Regulation of cytoplasmic mRNA decay. *Nat Rev Genet*, **13**, 246-259.

Schoning, J.C., Streitner, C., Meyer, I.M., Gao, Y. and Staiger, D. (2008) Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis. *Nucleic Acids Res*, **36**, 6977-6987.

Schwartz, S., Meshorer, E. and Ast, G. (2009) Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, **16**, 990-995.

Seo, P.J., Park, M.J., Lim, M.H., Kim, S.G., Lee, M., Baldwin, I.T., et al. (2012) A self-regulatory circuit of CIRCADIAN CLOCK-ASSOCIATED1 underlies the circadian clock regulation of temperature responses in Arabidopsis. *Plant Cell*, **24**, 2427-2442.

Severing, E.I., van Dijk, A.D., Stiekema, W.J. and van Ham, R.C. (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics*, **10**, 154.

Shen, Y., Zhou, Z., Wang, Z., Li, W., Fang, C., Wu, M., et al. (2014) Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell*, **26**, 996-1008.

Silverman, I.M., Li, F. and Gregory, B.D. (2013) Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. *Plant Sci*, **205-206**, 55-62.

Simpson, C.G., Jennings, S.N., Clark, G.P., Thow, G. and Brown, J.W. (2004) Dual functionality of a plant U-rich intronic sequence element. *Plant J*, **37**, 82-91.

Sorek, R., Ast, G. and Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res*, **12**, 1060-1067.

Staiger, D. and Brown, J.W. (2013) Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell*, **25**, 3640-3656.

Staudt, A.C. and Wenkel, S. (2011) Regulation of protein function by 'microProteins'. *EMBO Rep*, **12**, 35-42.

Stauffer, E., Westermann, A., Wagner, G. and Wachter, A. (2010) Polypyrimidine tract-binding protein homologues from Arabidopsis underlie regulatory circuits based on alternative splicing and downstream control. *Plant J*, **64**, 243-255.

Sterner, D.A., Carlo, T. and Berget, S.M. (1996) Architectural limits on split genes. *Proc Natl Acad Sci U S A*, **93**, 15081-15085.

Streitner, C., Simpson, C.G., Shaw, P., Danisman, S., Brown, J.W. and Staiger, D. (2013) Small changes in ambient temperature affect alternative splicing in *Arabidopsis thaliana*. *Plant Signal Behav*, **8**, e24638.

Su, Z., Wang, J., Yu, J., Huang, X. and Gu, X. (2006) Evolution of alternative splicing after gene duplication. *Genome Res*, **16**, 182-189.

Sugio, A., Dreos, R., Aparicio, F. and Maule, A.J. (2009) The cytosolic protein response as a subcomponent of the wider heat shock response in Arabidopsis. *Plant Cell*, **21**, 642-654.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A. and Brown, J.W. (2012) Alternative splicing in plants--coming of age. *Trends Plant Sci*, **17**, 616-623.

Syed, N.H., Prince, S.J., Mutava, R.N., Patil, G., Li, S., Chen, W., et al. (2015) Core clock, SUB1, and ABAR genes mediate flooding and drought responses via alternative splicing in soybean. *J Exp Bot*, **66**, 7129-7149.

Talerico, M. and Berget, S.M. (1994) Intron definition in splicing of small Drosophila introns. *Mol Cell Biol*, **14**, 3434-3445.

Tanabe, N., Yoshimura, K., Kimura, A., Yabuta, Y. and Shigeoka, S. (2007) Differential expression of alternatively spliced mRNAs of Arabidopsis SR protein homologs, atSR30 and atSR45a, in response to environmental stress. *Plant Cell Physiol*, **48**, 1036-1049.

Thatcher, S.R., Danilevskaya, O.N., Meng, X., Beatty, M., Zastrow-Hayes, G., Harris, C., et al. (2016) Genome-Wide Analysis of Alternative Splicing during Development and Drought Stress in Maize. *Plant Physiol*, **170**, 586-599.

Thomas, J., Palusa, S.G., Prasad, K.V., Ali, G.S., Surabhi, G.K., Ben-Hur, A., et al. (2012) Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of SCL33 pre-mRNA. *Plant J*, **72**, 935-946.

Tillemans, V., Dispa, L., Remacle, C., Collinge, M. and Motte, P. (2005) Functional distribution and dynamics of Arabidopsis SR splicing factors in living plant cells. *Plant J*, **41**, 567-582.

Tillemans, V., Leponce, I., Rausin, G., Dispa, L. and Motte, P. (2006) Insights into nuclear organization in plants as revealed by the dynamic distribution of Arabidopsis SR splicing factors. *Plant Cell*, **18**, 3218-3234.

Vogan, K.J., Underhill, D.A. and Gros, P. (1996) An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. *Mol Cell Biol*, **16**, 6677-6686.

Wachter, A., Ruhl, C. and Stauffer, E. (2012) The Role of Polypyrimidine Tract-Binding Proteins and Other hnRNP Proteins in Plant Splicing Regulation. *Front Plant Sci*, **3**, 81.

Wang, B.B. and Brendel, V. (2006) Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci U S A*, **103**, 7175-7180.

Wang, B.B., O'Toole, M., Brendel, V. and Young, N.D. (2008) Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. *BMC Plant Biol*, **8**, 17.

Warf, M.B. and Berglund, J.A. (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci*, **35**, 169-178.

Werneke, J.M., Chatfield, J.M. and Ogren, W.L. (1989) Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and Arabidopsis. *Plant Cell*, **1**, 815-825.

Wu, G. and Poethig, R.S. (2006) Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3. *Development*, **133**, 3539-3547.

Wu, H.P., Su, Y.S., Chen, H.C., Chen, Y.R., Wu, C.C., Lin, W.D., et al. (2014) Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in *Physcomitrella patens*. *Genome Biol*, **15**, R10.

Wu, J. and Baldwin, I.T. (2010) New insights into plant responses to the attack from insect herbivores. *Annu Rev Genet*, **44**, 1-24.

Wu, J., Hettenhausen, C., Meldau, S. and Baldwin, I.T. (2007a) Herbivory rapidly activates MAPK signaling in attacked and unattacked leaf regions but not between leaves of *Nicotiana attenuata*. *Plant Cell*, **19**, 1096-1122.

Wu, J., Kang, J.H., Hettenhausen, C. and Baldwin, I.T. (2007b) Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in *Nicotiana attenuata*. *Plant J*, **51**, 693-706.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (2006) Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu Rev Plant Biol*, **57**, 781-803.

Yan, K., Liu, P., Wu, C.A., Yang, G.D., Xu, R., Guo, Q.H., et al. (2012) Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in *Arabidopsis thaliana*. *Mol Cell*, **48**, 521-531.

Yang, G.D., Yan, K., Wu, B.J., Wang, Y.H., Gao, Y.X. and Zheng, C.C. (2012a) Genomewide analysis of intronic microRNAs in rice and Arabidopsis. *J Genet*, **91**, 313-324.

Yang, X., Zhang, H. and Li, L. (2012b) Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. *Plant J*, **70**, 421-431.

Yoshimura, K., Yabuta, Y., Ishikawa, T. and Shigeoka, S. (2002) Identification of a cis element for tissue-specific alternative splicing of chloroplast ascorbate peroxidase pre-mRNA in higher plants. *J Biol Chem*, **277**, 40623-40632.

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., et al. (2010) Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq. *Plant Physiol*, **152**, 1787-1795.

Zhang, F., Zhu, G., Du, L., Shang, X., Cheng, C., Yang, B., et al. (2016) Genetic regulation of salt stress tolerance revealed by RNA-Seq in cotton diploid wild species, *Gossypium davidsonii*. *Sci Rep*, **6**, 20582.

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., et al. (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*, **20**, 646-654.

Zhang, X.C. and Gassmann, W. (2003) RPS4-mediated disease resistance requires the combined presence of RPS4 transcripts with full-length and truncated open reading frames. *Plant Cell*, **15**, 2333-2342.

Zhang, X.C. and Gassmann, W. (2007) Alternative splicing and mRNA levels of the disease resistance gene RPS4 are induced during defense responses. *Plant Physiol*, **145**, 1577-1587.

Zhang, X.N. and Mount, S.M. (2009) Two alternatively spliced isoforms of the Arabidopsis SR45 protein have distinct roles during normal plant development. *Plant Physiol*, **150**, 1450-1458.

2. Overview of Manuscripts

Manuscript I

Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*

Zhihao Ling, Wenwu Zhou, Ian T. Baldwin and Shuqing Xu

Published in The Plant Journal (2015), 84, 228–243
Doi: 10.1111/tpj.12997.

In eukaryotic organisms, transcriptional regulations by both expression differences and alternative splicing (AS) are thought to be important for their responses to abiotic and biotic stresses. For plants, insect herbivores present one of the most important biotic stresses and their attack can elicit rapid changes in gene expression, changes which are essential for induced defenses that effectively reduce attack rates. However, insect herbivore-induced AS in plants remains unknown. To address this bias, we analyzed the genome-wide herbivore induced AS responses in both leaves and roots of *N. attenuata*, an ecological model plant, and investigated the possible signaling functions of genes involved in herbivore-induced AS machinery. The study reveals that an above-ground biotic stress can induce strong AS responses in a below-ground tissue. Furthermore, we demonstrated that the induced AS are precisely regulated and likely contribute to anti-herbivore defenses of plants.

ZL designed the research, performed the experiments, analyzed data and drafted the manuscript. WZ performed the experiments. ITB designed the research and helped to draft the manuscript. SX designed the research, analyzed data and drafted the manuscript.

Manuscript II

Wild tobacco genomes reveal the evolution of prolific nicotine production

Shuqing Xu, Thomas Brockmöller, Aura Navarro-Quezada, Heiner Kuhl, Klaus Gase, Zhihao Ling, Wenwu Zhou, Christoph Kreitzer, Mario Stanke, Haibao Tang, Eric Lyons, Priyanka Pandey, Shree P. Pandey, Bernd Timmermann, Emmanuel Gaquerel, and Ian T. Baldwin

Submitted to Nature Plants (10.2016)

In this manuscript, we sequenced assembled and analyzed the high-quality genomes of two wild tobaccos, *Nicotiana attenuata* and *N. obtusifolia* to investigate their adaptive traits in nature. Using phylogenomic analyses, we demonstrated that the pathway gradually evolved from two duplicated ancient primary metabolic pathways is followed by the rapid acquisition of root-specific gene expression. Furthermore, we revealed that in *Nicotiana* genomes, a rapidly expanding of transposable elements (TEs) insertions occurs after the Solanaceae whole genome triplication event, which contributed to the evolution of herbivory-induced signaling and defense, such as nicotine biosynthesis. These TE insertions also play roles in leading the expression divergences among duplicated genes. We also found TE insertions in regulatory regions of duplicated genes that incorporated transcription factor binding motifs are very likely contributing to the coordinated metabolic flux of the related biosynthetic pathway.

SX and ITB conceived and coordinated the project. 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. HK and SX assembled the genomes. 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. WZ, CK and KG validated promoter region of nicotine biosynthesis genes using Sanger sequencing. SX, TB, HT, MS and EL annotated protein coding genes in the genomes. PP and SPP annotated smRNAs in *N. attenuata*. SX, EG and ITB wrote the manuscript.

Manuscript III

Deep learning rapid evolution of alternative splicing in plants

Zhihao Ling, Thomas Brockmöller, Ian T. Baldwin and Shuqing Xu

To be submitted to Genome Biology

In this manuscript, we performed an intensive genome-wide transcriptome analysis of six plant species to investigate the evolution of alternative splicing (AS) in plant. We found that the global AS patterns are more similar among different tissues within same species than same tissue among different species, suggesting AS in plants evolved rapidly. Using a machine learning approach, we found the splicing codes in plants are largely conserved and the rapid divergence of determinant sequences of AS mainly contributes to the rapid turnover of AS between species. Furthermore, we found the group of AS which would generate transcripts with premature termination code (PTC), although only account for a minor portion of total AS, is more conserved than the other AS, suggesting an important regulation role of PTC+ AS at the post-transcriptional level in plants.

ZL designed the research, analyzed data and drafted the manuscript. TB identified one-to-one orthologous gene pairs and estimated the gene family size. ITB designed the research and helped to draft the manuscript. SX designed the research, analyzed data and drafted the manuscript.

3. Manuscripts

Manuscript I

Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*

Zhihao Ling, Wenwu Zhou, Ian T. Baldwin and Shuqing Xu

Published in The Plant Journal (2015), 84, 228–243

Doi: 10.1111/tpj.12997.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **84**, 228–243 doi: 10.1111/tpj.12997

RESOURCE

Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*

Zhihao Ling, Wenwu Zhou, Ian T. Baldwin and Shuqing Xu*

Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Hans-Knöll-Straße 8, D-07745 Jena, Germany

Received 8 June 2015; revised 4 August 2015; accepted 11 August 2015; published online 26 August 2015.
*For correspondence (e-mail sxu@ice.mpg.de).
Accession numbers: The RNA-seq data for generating the results in this study are deposited in NCBI short reads archive under the Bio Project ID (PRJNA223344), the sequences of genes that were selected for alternative splicing and expression analysis has been deposited at DDBJ/EMBL/GenBank under the accession GCZL01000000.

SUMMARY

Changes in gene expression and alternative splicing (AS) are involved in many responses to abiotic and biotic stresses in eukaryotic organisms. In response to attack and oviposition by insect herbivores, plants elicit rapid changes in gene expression which are essential for the activation of plant defenses; however, the herbivory-induced changes in AS remain unstudied. Using mRNA sequencing, we performed a genome-wide analysis on tobacco hornworm (*Manduca sexta*) feeding-induced AS in both leaves and roots of *Nicotiana attenuata*. Feeding by *M. sexta* for 5 h reduced total AS events by 7.3% in leaves but increased them in roots by 8.0% and significantly changed AS patterns in leaves and roots of existing AS genes. Feeding by *M. sexta* also resulted in increased (in roots) and decreased (in leaves) transcript levels of the serine/arginine-rich (SR) proteins that are involved in the AS machinery of plants and induced changes in SR gene expression that were jasmonic acid (JA)-independent in leaves but JA-dependent in roots. Changes in AS and gene expression elicited by *M. sexta* feeding were regulated independently in both tissues. This study provides genome-wide evidence that insect herbivory induces changes not only in the levels of gene expression but also in their splicing, which might contribute to defense against and/or tolerance of herbivory.

Keywords: alternative splicing, herbivore-induced responses, jasmonic acid, serine/arginine-rich (SR) proteins, *Nicotiana attenuata*, *Manduca sexta*.

INTRODUCTION

Alternative splicing (AS), a mechanism by which the same precursor mRNA (pre-mRNA) generates two or more different transcripts by using different splice sites, is widespread in higher eukaryotes (Black, 2003). For example, more than 95% of intron-containing genes in humans are alternatively spliced (Pan *et al.*, 2008). On the basis of expressed sequence tag data AS was found to be less common in plants (Iida *et al.*, 2004; Campbell *et al.*, 2006), but recent studies using next-generation sequencing (NGS) technologies have shown that up to 61% of intron-containing genes are alternatively spliced in *Arabidopsis thaliana* (Marquez *et al.*, 2012). More than 50% of multi-exon genes were found to be alternatively spliced in moss (*Physcomitrella patens*), suggesting that AS also occurs in non-vascular plants (Wu *et al.*, 2014). Alternatively spliced transcripts can be broadly classified into four basic types: intron retention (IR), alternative acceptor sites (AltA), alternative donor sites (AltD) and exon skipping (ES) (Black, 2003). While ES was found to be the most common type of AS in mammals (Pan *et al.*, 2008), IR is the most abundant type of AS in plants (Zhang *et al.*, 2010; Marquez *et al.*, 2012; Shen *et al.*, 2014; Thatcher *et al.*, 2014; Wu *et al.*, 2014).

Pre-mRNA splicing is catalyzed by the spliceosome, consisting of several uridylate-rich small nuclear RNAs (UsnRNAs) and more than 300 proteins (Jurica and Moore, 2003; Will and Luhrmann, 2011). The core of the spliceosome, which is largely conserved among plants and Metazoa, consists of five small nuclear ribonucleoproteins (snRNPs) and numerous non-snRNP proteins (Reddy, 2007). In plants, one highly conserved protein family, serine/arginine-rich (SR) proteins, are involved in regulating the choice of splice site by interacting with the spliceosome (Wang and Brendel, 2004; Isshiki *et al.*, 2006; Reddy, 2007; Reddy and Ali, 2011). The number of SR proteins varies

among different plant species, with 18 in *A. thaliana*, 22 in rice and 25 in soybean (Reddy and Ali, 2011). The different SR proteins can be classified into six subfamilies (SR, SC, RSZ, SCL, RS2Z and RS) based on phylogenic analysis (Barta *et al.*, 2010). Among these, three subfamilies (SR, SC and RSZ) also exist in mammals while the other three (SCL, RS2Z and RS) are plant-specific (Reddy and Ali, 2011). In addition, two SR-like genes, *SR-45* and *SR-45a*, that were not classified as SR genes according to the proposed criteria (Manley and Krainer, 2010) are also involved in regulating splice-site choice in plants (Ali *et al.*, 2007; Tanabe *et al.*, 2007; Cruz *et al.*, 2014). The SR and SR-like genes can be regulated by abiotic stresses at multiple levels (Duque, 2011), such as changes in expression (Filichkin *et al.*, 2010), splicing patterns (Lopato *et al.*, 1999; Kalyna *et al.*, 2003; Isshiki *et al.*, 2006), phosphorylation status and subcellular localization (Ali *et al.*, 2003), thus mediating stress-induced changes in AS. However, the mechanisms of stress-induced SR and SR-like gene regulation remain unclear.

Due to their sessile lifestyle, plants respond strongly to environmental stresses and AS is important for physiological responses and adaptations to different stresses (Mastrangelo *et al.*, 2012). A variety of stresses, such as cold (Leviatan *et al.*, 2013), light (Wu *et al.*, 2014), salt (Ding *et al.*, 2014) and pathogens (Lopato *et al.*, 1999; Howard *et al.*, 2013), can induce global changes in AS in plants, and changes in AS patterns have been found to function in adaptations to both biotic and abiotic stresses. For example, AS mediates the responses of the circadian clock to temperature changes (James *et al.*, 2012) and regulates the activation of the mitogen-activated protein kinase (MAPK) cascade, which in turn mediates different abiotic stress responses in *A. thaliana* (Lin *et al.*, 2010). In addition, intraspecific variation in AS of a proline synthetic enzyme, Δ1-pyrroline-5-carboxylate synthetase1 (P5CS1), was found to be correlated with drought-induced accumulation of proline and was under positive selection in *A. thaliana*, consistent with the hypothesis that AS of P5SC1 may mediate drought resistance (Kesari *et al.*, 2012). Furthermore, the AS in two resistance genes (*R*), which belong to the nucleotide-binding site (NBS) leucine-rich repeat region (LRR) class, are necessary to confer complete resistance to pathogens (Dinesh-Kumar and Baker, 2000; Weaver *et al.*, 2006).

Insect herbivory is a major biotic stress with large consequences for plant fitness (Kessler *et al.*, 2004; Maron and Crone, 2006; Wu and Baldwin, 2010). In response to herbivore attack, plants rapidly elicit changes in levels of phytohormone and gene expression which are essential for the physiological changes that mediate defense responses (Kessler *et al.*, 2004; Maron and Crone, 2006; Wu and Baldwin, 2010). Among these phytohormonal changes, rapid accumulation of jasmonic acid (JA) and its derivatives after herbivore attack play a central role in the activation of defense responses against many insect herbivores (Farmer and Ryan, 1990; Kessler *et al.*, 2004; Wu and Baldwin, 2010; Ballare, 2011). Furthermore, these responses are highly tissue specific, with some responses spreading rapidly throughout the plant (Erb *et al.*, 2009). For instance, the defense signal JA and other unknown mobile signals can either act locally or travel to systemic tissues (Wu and Baldwin, 2010) to fine-tune the transcriptome and metabolism of the whole plant to activate defense, avoidance or tolerance responses (Zhang and Baldwin, 1997; Gulati *et al.*, 2013; Fragoso *et al.*, 2014). Using reverse genetic approaches, many of these herbivory-induced transcriptomic and metabolomic changes have been demonstrated to be important for anti-herbivore defenses in plants (Kessler *et al.*, 2004; Kandoth *et al.*, 2007; Wu *et al.*, 2007; Meldau *et al.*, 2009; Dinh *et al.*, 2013). However, herbivory-induced AS responses in plants and their potential contributions to anti-herbivore defense and/or tolerance remain unknown.

In this study, we investigated herbivory-induced genome-wide AS in *Nicotiana attenuata*, an ecological model plant native to western North America. We sequenced the transcriptomes of both leaves and roots of *N. attenuata* control plants as well as plants that were attacked by *Manduca sexta*, a specialist herbivore of Solanaceae. We analyzed the AS responses induced by *M. sexta* feeding on *N. attenuata* at three different levels: (i) novel AS events, which were specifically induced or suppressed by *M. sexta* feeding; (ii) differentially spliced (DS) genes, defined as genes that were alternatively spliced in both control samples and those subject to *M. sexta* feeding but which show significantly different splicing ratios; and (iii) expression of genes involved in the splicing machinery. From both the changes in global AS responses and expression of SR genes, we conclude that *M. sexta* attack elicits AS in both leaves and roots, and the effects are greater in roots. Functional analyses of the genes involved in herbivory-induced AS responses indicate that these induced AS responses might contribute to anti-herbivore defenses. Furthermore, using JA-deficient plants, we provide evidence that herbivory-induced changes in SR gene expression in roots are JA dependent.

RESULTS

RNA sequencing, transcriptome assembly and AS detection

To investigate *M. sexta* feeding-induced transcriptomic responses in *N. attenuata* we sequenced 18 mRNA samples from the leaves (local tissue) and roots (systemic tissue) of control plants and plants that had been attacked by *M. sexta* larvae for 5 and 9 h, using three replicates for each time point and treatment. In total, we obtained about 722 million trimmed high-quality reads (on average about 40 million reads per sample) (see Table S1 in Supporting

Information). For each sample, at least 85% of the trimmed reads were uniquely mapped to the *N. attenuata* draft genome (v 1.0), and at least 22% of trimmed reads were uniquely mapped to splice junctions (SJs) (Table S1). The transcripts were assembled using all mapped reads. After filtering out low-expression transcripts (details in Experimental procedures), we obtained 34 359 high-quality transcripts. These transcripts captured 92% of the annotated transcripts (16 896 of 18 365 transcripts) that were known to be expressed in leaves and/or roots (estimated based on more than 350 and 360 million clean reads generated by RNA-sequencing of a single leaf and root, respectively). In addition, 15 994 transcripts (47% of all assembled transcripts) were identified as novel transcripts, of which the majority (86%) was located in known gene regions (Figure S1). The length of most assembled transcripts was between 1000 and 2500 bp, with an average length of 1806 bp (Figure S1), which is similar to the full cDNA length distribution in *A. thaliana* (Alexandrov *et al.*, 2006). Approximately 75% of total assembled transcripts and 72% of assembled novel transcripts had a complete open reading frame (ORF), indicating that the majority of assembled transcripts can be translated into proteins. Furthermore, about 16.4% of expressed novel assembled transcripts contained premature termination codons (PTC), suggesting that most of the novel transcripts are not the target of nonsense-mediated mRNA decay (NMD) (Chang *et al.*, 2007).

To evaluate whether the sequencing depth was sufficient to discover the majority of the transcripts and exon exon junctions, which is of central importance for the downstream AS analysis, we subsampled different numbers of mapped reads and calculated the saturation curve for both transcript assembly and junction detection (details in Experimental procedures). The analysis showed that at a sequencing depth greater than 320 million uniquely mapped reads (about 17.8 million reads per sample), the percentage coverage of novel transcripts and junctions only increases marginally with higher sequencing depth (Figure S2). This result suggests that our sequencing depth (about 620 million uniquely mapped reads) was sufficient to assemble the majority of novel transcripts and discover most of the exon exon junctions.

To detect AS events in *N. attenuata*, we first extracted the coordinates of all SJs and retained only those with overhangs longer than 13 bp (see details in Experimental procedures). In total, 153 824 high-confidence SJs were retained for the downstream analysis. Based on the mapping information of these SJs among all 18 samples, we identified 50 932 AS events in 12 967 genes, which revealed that about 66% of the intron-containing genes expressed in the leaves or roots of *N. attenuata* were alternatively spliced. Similar results were found when we analyzed the control samples and those subject to *M. sexta* feeding separately (Table S2). Among all AS events, intron retention (IR) was the most abundant AS type in *N. attenuata*, followed by alternative 3′ accepter site (AltA), alternative 5′ donor site (AltD) and exon skipping (ES) (Figure S3). This is similar to the global AS pattern of *A. thaliana* (Marquez *et al.*, 2012), *Physcomitrella patens* (Wu *et al.*, 2014) and *Glycine max* (Shen *et al.*, 2014), and suggests that IR is the most common type of AS among different plant families.

Attack by *M. sexta* induces changes in the total number of AS events in leaves and roots

We investigated insect herbivory-induced AS in *N. attenuata* leaves and roots separately by comparing the AS events between control and attacked plants. To reduce bias from unequal sequencing depth among samples, we subsampled 23 million (the smallest number of uniquely mapped reads among all samples) uniquely mapped reads from each sample. Unless specified, all of the subsequent comparative analyses were performed on this subsampled dataset. Interestingly, after *M. sexta* attack, the total number of AS events decreased in leaves but increased in roots. In leaves, the average number of total AS events decreased by 7.3% after 5 h of *M. sexta* feeding (control samples, 10 990; *M. sexta* feeding samples, 10 189). The decrease in AS in leaves was largely due to reductions in two types of AS, AltD and AltA, which were reduced by 10.2 and 8.2%, respectively, in comparison with those of control samples (Figure 1). In roots, 5 h of *M. sexta* feeding increased the total number of AS events by about 8.0% in comparison with control samples (control samples, 14 208; *M. sexta* feeding samples, 15 345). This increase was mainly due to increases in three types of AS events, AltD, AltA and IR, which showed respective increases of 5.6, 4.2 and 15.1% increases in comparison with controls (Figure 1). The pattern remained when only AS events that were shared among all three biological replicates were considered (Figure S4).

Previous studies have demonstrated that simulated *M. sexta* feeding induces large-scale changes in gene expression (Gilardoni *et al.*, 2010; Gulati *et al.*, 2014), and that changes in expression levels increase the likelihood of AS for particular genes (Shen *et al.*, 2014). Therefore, we reasoned that the increase and decrease of AS events could reflect changes in gene expression or gene splicing regulation, or both. To test whether attack by *M. sexta* induced novel AS events without regulating expression, we selected a subset of AS events that had similar read coverage for related regions between attacked and control samples (Figure S5). Since the expression levels of these junction regions were the same, the novel AS events identified from these subsets could only be due to AS regulation. The results based on this subset of AS events were consistent with the observation using all junction data (Figure S6). Furthermore, the number of AS events in both

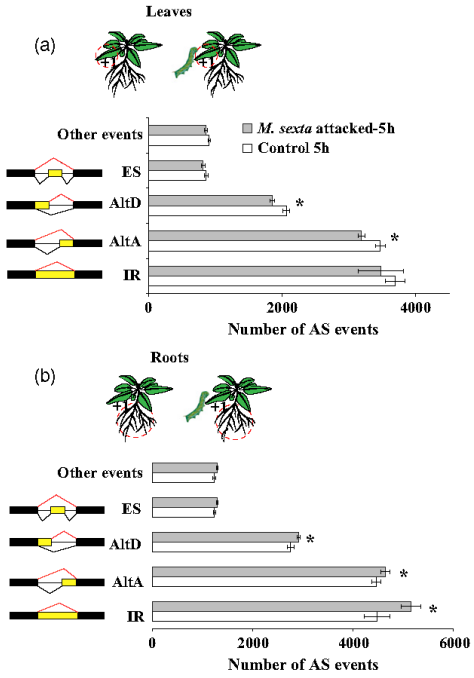


Figure 1. Total number of alternative splicing (AS) events in roots and leaves of *Nicotiana attenuata*.
The number of different types of AS events in leaves (a) and roots (b) of *N. attenuata* that were either not attacked or attacked for 5 h by *Manduca sexta* larvae (ES, exon skipping; AltD, alternative donor; AltA, alternative acceptor; IR, intron retention). Asterisks indicate the significance as determined by Student's *t*-test ($P < 0.05$). Error bars refer to standard errors.

leaves and roots showed no significant differences between samples that were attacked by *M. sexta* larvae for 5 or 9 h (Figure S6), although it was slightly increased in the leaves that were attacked for 9 h. These data show that feeding by *M. sexta* can elicit AS responses in genes without regulating their expression levels.

To investigate the potential biological functions of the AS events in leaves and roots regulated by *M. sexta* attack we performed Gene Ontology (GO) enrichment analysis. To avoid the effects caused by induced expression changes, we only focused on the junction regions that had similar expression levels in both control plants and those attacked by *M. sexta*. In leaves, we identified 678 control-specific AS events (suppressed by *M. sexta* attack) from 398 genes (Data S1) and these were enriched in four GO terms: 'phospholipid biosynthetic process', 'starch metabolic process', 'carotenoid biosynthetic process' and 'thylakoid membrane organization' (Figure 2). Among genes in the starch metabolic process, a β-amylase (*NaBAM2*) that interacts with starch or other α-1,4-glucans (Fulton *et al.*, 2008) had two transcripts, a dominant transcript (*NaBAM2.t1*) that was present in both attacked and control samples and a minor transcript (*NaBAM2.t2*) that was only found in control samples (Figure 2), indicating that the latter transcript was suppressed by *M. sexta* attack. The minor transcript was generated by an AS event found at an alternative 5′ donor site (AltD) on the second exon. Although both transcripts encode protein sequences with two domains the sugar-binding domain (Glyco_hydro_2_N) and the starch-binding domain (CBM_20) the minor transcript is shorter and lacks some amino acid residues at the amino-terminus. In addition to AS events that were suppressed in leaves, we also identified 398 AS events from 244 genes that were specifically elicited in leaves (Data S1). These genes were significantly enriched in the function '*O*-methyltransferase activity'.

In roots, while 698 AS events from 387 genes enriched in the 'acetate-CoA ligase activity' category were identified as being found only in controls (suppressed by *M. sexta* attack), 1011 AS events from 597 genes were found only in attacked root samples (Data S1). The genes involved in these *M. sexta* feeding-induced AS events were enriched in the functional categories of 'mRNA metabolic process', 'RNA splicing', 'lipid modification' and 'protein glycosylation', indicating that *M. sexta* attack induces changes in AS of genes involved in mRNA processing machinery in roots (Figure 2). In the functional group 'RNA splicing', a RNA splicing factor-like gene (de la Fuente van Bentem *et al.*, 2006), *NaESP3*, had a novel AltD splicing event in its second exon that only was found in the attacked samples. In comparison with its dominant transcript (*NaESP3.t1*), this novel splicing event resulted in a transcript (*NaESP3.t2*) that encodes a putatively fully functional protein but lacks the PWI domain: a domain of unknown function conserved in splicing factors (Figure 2). The two transcripts both had a complete ORF but differed in their functional domains, indicating that they might both be functional but involved in different biological processes.

In summary, our analysis showed that feeding by *M. sexta* elicits changes in AS in both leaves and roots of *N. attenuata*, with stronger effects in roots than in leaves. While feeding by *M. sexta* suppresses AS in genes involved in primary metabolisms in leaves, it elicits AS in genes involved in the mRNA processing and modification machinery of roots.

Attack by *M. sexta* affects the splicing ratio of existing AS events in leaves and roots

Splicing regulations can not only result in novel AS events but may also lead to changes in the relative expression ratio between alternatively spliced and dominant transcripts, which can be expressed by the percentage of splicing index

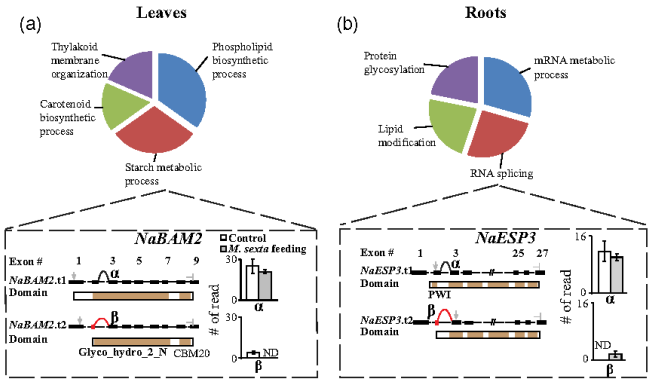


Figure 2. The enriched Gene Ontology (GO) terms of condition-specific alternative splicing (AS) events.
(a), (b) Enriched GO terms of novel AS events that were specifically suppressed in leaves (a) and induced in roots (b) after 5 h of feeding by *Manduca sexta*. The insert in each panel depicts one AS example from genes involved in the 'starch metabolic process' (left) and 'RNA splicing' (right). For each AS example, the resultant protein domain (filled in brown) and the number of reads supporting the dominant splicing (α) and minor splicing (β) events are shown. *NaBAM2*, a beta-amylase that degrades starch or other alpha-1,4-glucans, contains two domains: a sugar-binding domain (Glyco_hydro_2_N) and a starch-binding domain (CBM_20); *NaESP3*, a RNA splicing factor-like gene, contains six domains in its dominant transcript. The β-splicing event of the minor transcript results in the loss of the PWI domain, which is conserved in splicing factors with unknown function, and was only found in roots of plants attacked by *M. sexta*. The start and stop codons of each transcript are depicted as arrow and stop signs, respectively. ND refers to not detected.

(PSI). Using principal components analysis (PCA) based on the PSI values, we analyzed the global PSI changes in both leaves and roots. After 5 h of attack, both leaves and roots samples were different from controls, despite large variations among biological replicates (Figure S7a). We further identified DS genes by comparing the PSI of an AS event between attacked and control samples. In total, we found 180 (286 AS events) and 356 (557 AS events) DS genes (Data S2) after 5 h of *M. sexta* feeding in leaves and roots, respectively. These DS genes represent 2.3% (leaves) and 3.5% (roots) of all existing AS genes. The greater proportion of DS genes found in roots indicates that the contribution of herbivory-induced AS might be more pronounced in roots than in leaves, which is consistent with the pattern of the total number of novel AS events.

To validate the identified induced DS events, we measured the relative expression level of the transcripts from five selected candidate genes (three from leaves and two from roots), using quantitative PCR (qPCR) with transcript-specific primers. The RNA sequencing (RNA-seq) data showed that the relative expression of the minor transcripts was significantly increased in four genes and decreased in one gene (*NaEPSP1* in leaves) after 5 h of *M. sexta* attack (Figure 3). All these predictions were confirmed by our qPCR results (Figure 3, Table S3).

We further performed GO enrichment analysis to investigate the potential function of DS genes induced by *M. sexta* feeding. The DS genes in leaves were enriched in the 'nucleotide kinase activity', 'exopeptidase activity' and 'cellular biogenic amine biosynthetic process' functional categories, whereas DS genes in roots were enriched in the 'cysteine biosynthetic process' function.

Attack by *M. sexta* regulates the expression of SR genes

Genes of the SR protein family are involved in the regulation of AS in plants (Lopato *et al.*, 1999; Kalyna *et al.*, 2003; Isshiki *et al.*, 2006). In *N. attenuata*, we identified 20 SR and three SR-like genes from our current annotated draft genome (Figure S8). Based on their sequence similarity to the genes in Arabidopsis, we classified them into six SR subfamilies plus one SR-like subfamily (Duque, 2011; Cruz *et al.*, 2014). The expression levels of all identified SR genes were compared in attacked and control samples using the RNA-seq data. In leaves, 5 h of *M. sexta* feeding downregulated the expression of two SR genes (*NaRSZ22-1* and *NaSCL28* from the RSZ and SCL subfamilies, respectively) but did not upregulate any SR genes. In roots, 5 h of *M. sexta* feeding upregulated seven SR genes (*NaSR34a-2*, *NaRSZ21*, *NaSC35*, *NaSCL30a-1*, *NaSCL33*, *NaRS31-1* and *NaSR45a*-1) from six different subfamilies and downregulated three SR genes (*NaSR34-1/2* and *NaSCL30a-2* from the SR and SCL subfamilies, respectively; see Table S4). To further validate these changes, we measured the relative transcript accumulation of five selected SR and SR-like genes (one in leaves and four in roots) using qPCR in both attacked and control samples. The qPCR results were consistent with the predictions based on RNA-seq data for all measured genes (Figure S9). In addition, we also found that although the majority of SR genes underwent intensive AS (Table S5), only two AS

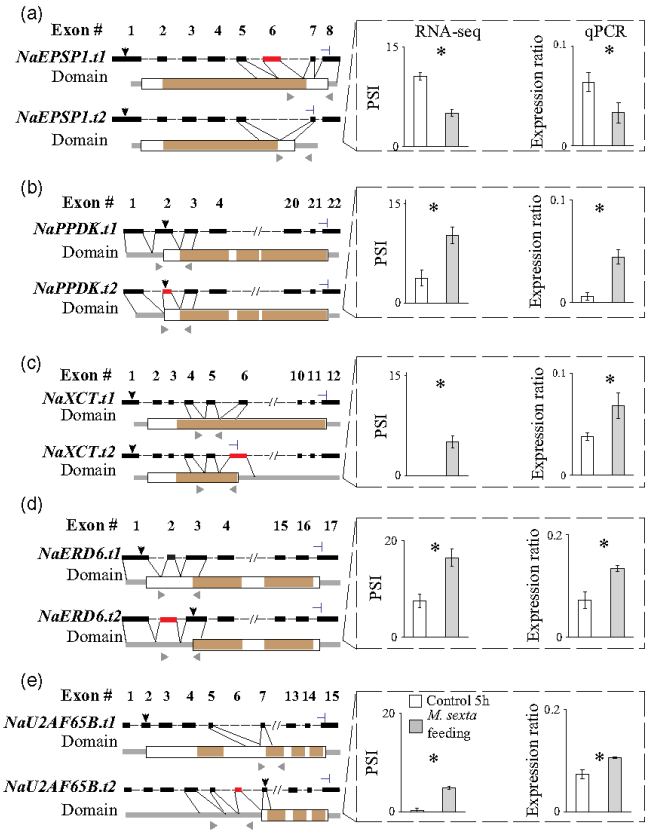


Figure 3. Validation of five differentially spliced (DS) genes induced by *Manduca sexta* attack.
Each panel depicts the identified alternative splicing (AS) events, the resulting protein domains (filled rectangle), the untranslated regions (UTRs; gray lines) and the positions for transcript-specific primers (gray arrows) for quantitative PCR. For each gene, the first (t1) and second (t2) transcripts indicate the dominant and minor transcript, respectively. The insert of each panel depicts the percentage splicing index (PSI) predicted from RNA-sequencing data and the expression ratio was calculated as the ratio between the expression of the minor transcript and total expression of both transcripts (t1 + t2). Exons involved in condition-specific AS events are indicated with different colors/shades. Gray bars refer to *M. sexta* feeding samples and white bars refer to control samples.
(a) Alternative splicing of *NaEPSP1* in leaves. *NaEPSP1* is a homolog of 5-enolpyruvylshikimate-3-phosphate synthase, a target of the herbicide glyphosate, and is associated with glyphosate tolerance.
(b) Alternative splicing of *NaPPDK* in leaves. *NaPPDK* is a homolog of pyruvate orthophosphate dikinase.
(c) Alternative splicing of *NaXCT* a leaves. *NaXCT* is a homolog of XAP5 CIRCADIAN TIMEKEEPER in Arabidopsis, which plays an important role in light regulation of the circadian clock, photomorphogenesis and ethylene regulation.
(d) Alternative splicing of *NaERD6* in roots. *NaERD6* is a homolog of a sugar transporter involved in early dehydration responses.
(e) Alternative splicing of *NaU2AF65B* in roots. *NaU2AF65B* is a homolog of *AtU2AF65B*, which encodes a large subunit of the splicing factor U2af. The start and stop codons are indicated with black arrows and stop signs, respectively, and the gray arrows indicate the positions of the primers used for quantitative PCR. An asterisk indicates a statistically significant difference as determined by Student's *t*-test ($P < 0.05$). Error bars refer to standard errors (SE).

events in *NaSCL30a-2* and *NaRS2Z32-1* were significantly regulated in roots by attack (Table S5) and none were regulated in leaves. These results suggest that *M. sexta* attack induces the expression changes of SR genes in both leaves and roots, but with stronger effects in roots than in leaves, again consistent with our observation on the differences in global changes in AS.

In plants, JA and its derivatives are the key phytohormone signals that mediate herbivory-induced transcriptome changes and defense responses (Kessler *et al.*, 2004; Kazan and Manners, 2008; Yan *et al.*, 2013). To investigate the role of JA in the regulation of *M. sexta* feeding-induced changes in SR gene expression, we measured the expression levels of the five induced candidate SR genes in a

JA-deficient line (ir*AOC*), in which a key JA biosynthesis gene, allene oxide cyclase (*AOC*), was silenced by RNA interference (RNAi) (Kallenbach *et al.*, 2012). In leaves, the expression of the SR gene *NaRSZ22*-1 significantly decreased after *M. sexta* feeding in both wild-type (WT) and JA-deficient plants (Figure 4), indicating that the downregulation of *NaRSZ22-1* in leaves is likely to be independent of JA. Interestingly, in roots, feeding by *M. sexta* resulted in increased expression of the four selected SR genes that were upregulated after feeding by *M. sexta* in WT but not JA-deficient plants, indicating that JA is required for herbivory-induced SR gene expression changes in *N. attenuata* roots (Figure 4).

We further investigated the role of JA in herbivory-induced changes in SR gene expression in roots by supplementing ir*AOC* plants with methyl jasmonate (MeJA) to increase the endogenous amount of free JA (Baldwin *et al.*, 1996; Wu *et al.*, 2008) and restore the plants' JA signaling. With the exception of one SR gene, *NaRSZ21*, which showed very low expression in ir*AOC* plants, even after MeJA supplementation, JA was involved in the upregulation of all three other SR genes. Although *NaSR45a-1* showed no significant changes in the roots of both control and MeJA supplemented ir*AOC* plants after *M. sexta* feeding, the overall expression levels of *NaSR45a-1* were much higher with MeJA supplementation (Figure 4c), indicating that upregulation of *NaSR45a-1* might be directly induced by accumulation of endogenous JA. Interestingly, while MeJA supplementation itself did not induce the expression of *NaSC35* and *NaSCL33*, increased levels of endogenous JA restored *M. sexta* attack-induced upregulation of *NaSC35* and *NaSCL33* in the roots of ir*AOC* plants (Figure 4d, e), suggesting that JA is required for their herbivory-induced upregulation. These results suggest that *M. sexta* feeding-induced changes in SR gene expression were probably JA independent in leaves but JA dependent in roots.

Attack by *M. sexta* regulates the AS of genes in the JA signaling pathway

Herbivory can rapidly increase transcript accumulation of plant genes involved in JA biosynthesis and JA-associated defense signaling (Reymond *et al.*, 2004; Chung *et al.*, 2008; Kim *et al.*, 2011). However, whether herbivory can also induce changes in AS in these genes remains unknown. We analyzed *M. sexta* feeding-induced AS responses in 14 genes that are known to be involved in JA biosynthesis and JA-mediated defense signaling. In leaves, six genes had at least two transcripts, suggesting that these genes were alternatively spliced, and 5 h of *M. sexta* attack upregulated 13 transcripts from eight genes and one novel AS event in jasmonoyl-L-isoleucine hydrolase 1 (*NaJIH1*) (Figure 5). The induced novel AS in *NaJIH1* was an ES event which took place at the fourth exon to produce two different transcripts. In comparison with the most abundant transcript (*NaJIH1.t1*), the protein-coding sequence of the minor transcript (*NaJIH1.t2*) is 35 amino acids (aa) shorter but contains a complete ORF (Figure 5b). From the RNA-seq data, the ES event was not detected in the control samples (PSI = 0), whereas *M. sexta* feeding increased its PSI to 2.5 (Figure 5c). To validate this change, we measured the relative expression of the two transcripts by qPCR with transcript-specific primers. Although both transcripts were detected in both control and attacked samples, the expression ratio of the minor transcript (with the ES event) increased from about 3.0% to about 7.5% (Figure 5d), consistent with the change found from the RNA-seq data. We further quantified expression changes of the two *NaJIH1* transcripts in ir*AOC* plants with and without MeJA supplementation. The dominant transcript

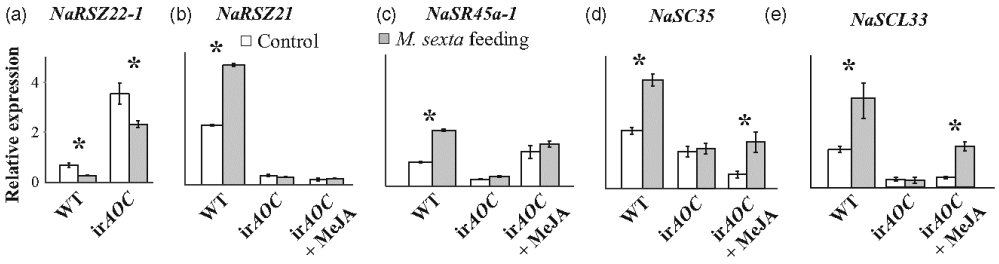


Figure 4. *Manduca sexta* attack-induced changes in SR gene expression in wild-type (WT) and jasmonic acid (JA)-deficient plants.
(a) Relative expression level of *NaRSZ22-1* in the leaves of WT and JA-deficient (ir*AOC*) plants.
(b)–(e) Relative expression level of *NaRSZ21* (b), *NaSR45a-1* (c), *NaSC35* (d) and *NaSCL33* (e) in the roots of WT and ir*AOC* plants. The ir*AOC* plants they were supplied with lanolin (ir*AOC*) or methyl jasmonate-containing lanolin (ir*AOC*+MeJA) for 24 h. White and gray bars refer to control samples and those subject to 5 h of *Manduca sexta* feeding, respectively. The expression levels of *NaRSZ21*, *NaSC35* and *NaSR45a-1* in the leaves and *NaRSZ22-1* in the roots of WT plants were not induced by *M. sexta* attack (Table S4). The asterisk indicates the significance determined by Student's *t*-test ($P < 0.05$). The error bars refer to standard error (SE).

(*NaJIH1.t1*) was not induced by *M. sexta* feeding in ir*AOC* plants but was in the ir*AOC* plants that received MeJA supplementation. This demonstrates that JA was required for *M. sexta* feeding-induced upregulation of the *NaJIH1* dominant transcript. Interestingly, *M. sexta* feeding did not induce accumulations of the minor transcript (*NaJIH1.t2*) in ir*AOC* plants, either with or without MeJA supplementation, but only in WT plants (Figure 5d). However, MeJA supplementation increased both the expression level and the relative ratio of *NaJIH1.t2* in ir*AOC* plants even without *M. sexta* feeding (Figure 5d), suggesting that the upregulation of *NaJIH1.t2* might only be regulated by increased endogenous JA levels.

In roots, 8 of 14 JA signaling genes had at least two transcripts, which resulted in a total of 19 different transcripts. Although *M. sexta* feeding did not upregulate any of these transcripts, it suppressed an AS event in *NaJAZe* (Figure S10). This *M. sexta*-suppressed AS event was an alternative first exon, which generated two different transcripts starting with a different first exon: one dominant transcript (*NaJAZe.t1*) and a minor transcript (*NaJAZe.t2*) which was 30 aa shorter but contained a complete ORF (Figure S10). From the RNA-seq data, 5 h of *M. sexta* attack reduced the PSI value of the minor transcript from about 6.3% to 0 (Figure S10). We conclude that *M. sexta* feeding can induce AS in JA signaling genes in both leaves and roots.

Changes of gene expression and splicing ratio induced by *M. sexta* attack are separately regulated

To investigate whether these attack-induced responses in transcript splicing and gene expression are independently regulated, we identified *M. sexta* 5-h feeding-induced differentially expressed (DE) genes and compared them

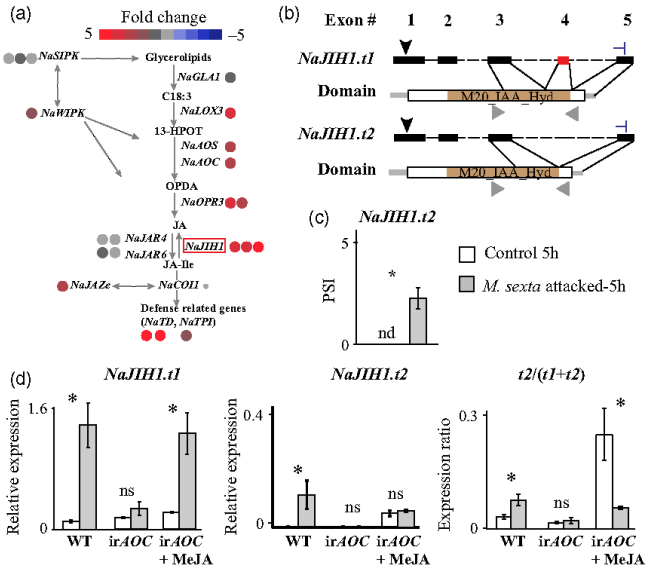


Figure 5. Alternative splicing (AS) events in genes of the jasmonic acid (JA) signaling pathway in *Nicotiana attenuata* leaves induced by *Manduca sexta* attack.
(a) Expression changes for the transcripts of JA signaling genes induced by *M. sexta* attack. Each filled circle refers to one transcript with the color indicating the $\log_2$ fold change in its abundance in comparison with control samples. The red rectangular box depicts the gene that contains the differentially spliced (DS) event induced by *M. sexta* attack. *NaJIH1*, jasmonoyl-l-isoleucine hydrolase 1; *NaSIPK*, salicylate-induced protein kinase; *NaWIPK*, wound-induced protein kinase; *NaGLA1*, glycerolipase 1; *NaLOX3*, lipoxygenase 3; *NaAOS*, allene oxide synthase; *NaAOC*, allene oxide cyclase; *NaOPR3*, OPDA reductase 3; *NaJAR4/6*, jasmonate-resistant 4/6; *NaCOI1*, coronatine-insensitive protein 1; *NaJAZe*, jasmonate ZIM-domain protein; *NaTD*, threonine deaminase; *NaTPI*, trypsin proteinase inhibitor.
(b) The *M. sexta* attack induced a DS alternative splicing (AS) event on *NaJIH1*. The lines (untranslated regions) and boxes (exons) refer to the gene structure, and the filled brown color refers to the predicted protein domains. The start and stop codons are indicated with a black arrow and stop sign, respectively. The exon involved in the AS event is indicated in red, and the primers used for quantitative (q)PCR are indicated with gray arrows.
(c) The percentage of splicing index (PSI) of the minor transcript resulting from the *M. sexta* attack-induced AS event. The PSI was predicted from RNA sequencing data.
(d) Transcript abundance of both minor (left, *NaJIH1.t1*) and dominant (middle, *NaJIH1.t1*) transcripts resulting from the *M. sexta* attack-induced AS event. The right panel shows the expression ratio calculated based on qPCR data in both wild-type (WT) and ir*AOC* plants [ir*AOC* plants, they were supplied with lanolin (ir*AOC*) or methyl jasmonate-containing lanolin (ir*AOC*+MeJA) for 24 h]. The asterisk indicates a significant difference determined by Student's *t*-test ($P < 0.05$). Error bars depict standard error (SE).

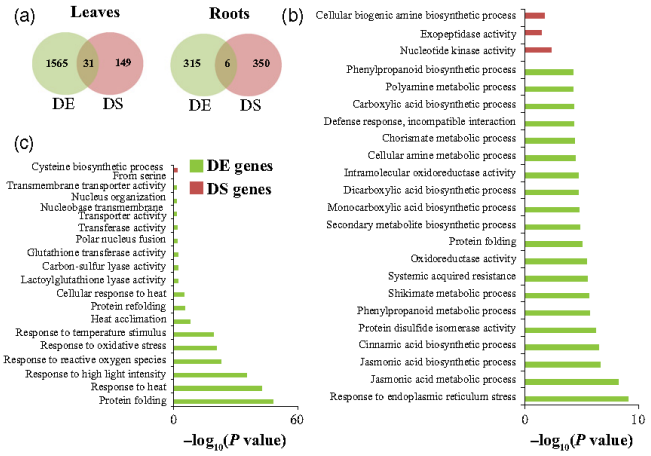


Figure 6. There is little overlap between *Manduca sexta* attack-induced gene alternative splicing and expression changes in *Nicotiana attenuata*.
(a) *Manduca sexta* attack induced differentially expressed (DE) and differentially spliced (DS) genes in both leaves (left) and roots (right).
(b), (c) Overlaps of the Gene Ontology (GO) terms enriched in DE and DS genes in leaves (b) and roots (c). The DS and DE genes are shown in different colors/shades.

with DS genes in roots and leaves. In leaves, 1596 DE genes and 180 DS genes were identified (Figure 6), whereas in roots, 321 DE genes and 356 DS genes were identified (Figure 6). Interestingly, only 31 (leaves) and 6 (roots) genes were found to be both differentially expressed and spliced (Figure 6), suggesting little overlap between DE and DS genes. We further performed GO enrichment analysis on the *M. sexta* feeding-induced DE and DS genes in leaves and roots. Overall, although the DE genes were enriched in more GO terms than the DS genes in both leaves and roots, little overlap was found. In leaves, while DE genes were enriched in 'jasmonic acid metabolic process', 'secondary metabolite biosynthetic process', 'defense responses' etc., DS genes were enriched in 'nucleotide kinase activity', 'exopeptidase activity' and 'cellular biogenic amine biosynthetic process' (Figure 6). Similarly, the DE and DS genes in roots were also enriched in different GO terms (Figure 6). Furthermore, we performed GO enrichment analysis for the up- and downregulated DE genes separately. Most of the DE genes in leaves were upregulated (1225 genes) and were similarly enriched in GO terms with those of the DE genes (Figure S11). For downregulated genes, several new enriched GO terms were found such as 'shoot system morphogenesis', 'photosynthesis' and 'starch metabolic process' (Figure S11). However, none of these overlapped with the GO terms that DS genes of leaves were enriched in. In roots, while no enriched GO term was found in downregulated genes, the enriched GO terms for upregulated DE genes, which made up the majority of DE genes (Figure S11), also showed no overlap with the enriched GO terms in DS genes. Taken together, our results showed that *M. sexta* feeding-induced AS and gene expression responses are two parallel processes.

DISCUSSION

Insect herbivory can induce rapid defense responses in plants at both phytohormone and transcriptomic levels (Wu and Baldwin, 2010; Gulati *et al.*, 2013, 2014). However, most of our current knowledge on herbivory-induced transcriptomic defense responses has been focused on changes in gene expression. Here, using RNA-seq, we provide evidence that insect herbivory can induce genome-wide AS responses in both leaves (local tissue) and roots (systemic tissue) and that the effects are stronger in roots. Functional analyses of these induced AS responses suggest that induced AS responses in both leaves and roots might contribute to anti-herbivore defenses and/or tolerance. Furthermore, our data also suggest that the induced AS responses in roots are likely regulated through the expression changes in SR genes. Using transgenic plants, we demonstrated that JA is involved in herbivory-induced upregulation of SR genes that are likely associated with AS responses.

Manduca sexta attack-induced AS responses in leaves and roots may contribute to anti-herbivore defenses and/or tolerance

Attack by *M. sexta* elicits changes in the total number of novel AS events, the ratio of AS to existing AS genes and the expression of SR genes in *N. attenuata*. These data suggest that in addition to previously described responses at the level of gene expression (Gulati *et al.*, 2013, 2014) insect herbivory can also induce transcriptomic responses at the AS level. Interestingly, the *M. sexta*-attack induced AS responses are more pronounced in roots than leaves (Figures 1 and 6), which is opposite to the responses at the gene expression level. This indicates that the elicitation of

changes in gene expression and AS are two parallel processes that function independently. Indeed, the comparison between herbivory-induced DS and DE genes showed that there is little overlap between them (Figure 6), which provides further support for the responses at the AS and expression levels being independently regulated. In *A. thaliana*, a similar pattern was also found when plant roots were stressed by Fe deficiency (Li *et al.*, 2013). Although additional comparisons between stress-induced gene expression and AS responses in different plant species under different stresses are needed to draw strong conclusions, this analysis, together with previous studies, suggests that the parallel regulations of AS and expression induced by stresses might be common in plants.

The biological functions and significance of stress-induced AS responses are often debated (Melamud and Moult, 2009; Zhang *et al.*, 2009, 2015; Li *et al.*, 2013): are they merely experimental artifacts/transcriptional noise or do they regulate and contribute to the physiological adaptation to stresses? The results of this analysis are consistent with the adaptive scenario, for the following reasons. First, a high confirmation rate (100%) for identified DS genes and changes in SR gene expression using qPCR suggests the *M. sexta* attack-induced AS responses are likely not experimental artifacts. Second, a recent study suggested that the mRNA extraction method might isolate both mature and incompletely spliced pre-mRNAs (Zhang *et al.*, 2015), thus leading to over-estimation of the abundance of AS. To investigate this issue, we calculated the percentage of novel transcripts that contain premature termination codons (PTC) that are not translated into proteins and degraded by the NMD pathway (Chang *et al.*, 2007). If the induced novel AS events had resulted from incompletely spliced pre-mRNAs, they would contain a higher proportion of PTC-containing transcripts (which would lead to unproductive mRNAs) than the genome-wide average. However, the portion of PTC-containing transcripts generated by *M. sexta* attack is only about 11% (38 in 338), which is even lower ($P = 0.03$, Fisher's exact test) than the overall proportion of PTC-containing transcripts in expressed novel transcripts (about 16.4%, 2118 in 12 930). Although a higher proportion of PTC-containing transcripts does not necessarily mean that they originated from incompletely spliced premature mRNA or other artifacts, a lower proportion of PTC-containing transcripts suggests that the incompletely spliced pre-mRNAs explanation is less likely.

Third, the AS events that were specifically suppressed in leaves and induced in roots were found to be enriched in specific gene ontologies (Figure 2), which might directly or indirectly contribute to anti-herbivore defenses and/or tolerance. In leaves, *M. sexta* attack suppresses AS in genes involved in primary metabolism, such as 'starch metabolic process' (Figures 2 and S11). The regulation of primary metabolism after herbivore attack was found to be common in plants (Schwachtje and Baldwin, 2008; Bilgin *et al.*, 2010), and can directly or indirectly contribute to anti-herbivore resistance (Mitra and Baldwin, 2008, 2014; Schwachtje and Baldwin, 2008). In roots, the *M. sexta*-induced novel AS events were enriched in functions related to RNA modification and protein glycosylation (Figure 2), indicating that the induced AS events may be involved in post-transcriptional regulation and transcriptomic fine-tuning in *N. attenuata* roots. In *Nicotiana*, the root plays a central role in anti-herbivore defenses and tolerance. For example, the biosynthesis of nicotine, which is one of the major anti-herbivore defense metabolites (Steppuhn *et al.*, 2004), takes place in roots. In addition, the roots also play a key role in resource reallocation and the ability for regrowth (Schwachtje *et al.*, 2006; Machado *et al.*, 2013; Vriet *et al.*, 2014). We hypothesize that *M. sexta* attack-induced AS responses in *N. attenuata* roots might contribute to the regulation of herbivory-induced defenses and/or tolerance (Steppuhn and Baldwin, 2007; Erb *et al.*, 2009). Further investigations that specifically manipulate the herbivory-induced AS responses in *N. attenuata* roots are required to test this hypothesis.

Finally, our analysis revealed that, in leaves, herbivory can induce AS events in genes involved in JA signaling in both leaves (*NaJIH1*) and roots (*NaJAZe*). NaJAZe is a member of the JASMONATE ZIM DOMAIN (JAZ) protein family, which functions as a negative regulator of jasmonic acid signaling in plants. NaJIH1 was shown to reduce levels of jasmonoyl-L-isoleucine and attenuate the herbivory-induced defense responses (Woldemariam *et al.*, 2012). Furthermore, our results also showed that herbivory likely regulates the AS of *NaJIH1* through increases in endogenous JA levels (Figure 5). This indicates that the regulated AS event in *NaJIH1* might be involved in the feedback loop between perception and metabolism of JA and thus may help to sustain JA signaling in leaves.

In summary, this analysis suggests that the herbivory-induced AS responses in *Nicotiana* plants are another important means of fine-tuning the transcriptome which is independent of changes in gene expression. The induced AS responses likely result from precise transcriptomic regulations and may contribute to anti-herbivore defenses and/or tolerance.

Alternative splicing responses in roots induced by *M. sexta* attack are associated with expression changes in the SR genes

In plants SR proteins can interact with the spliceosome to generate splicing responses (Lazar *et al.*, 1995; Golovkin and Reddy, 1998, 1999; Reddy, 2004; Reddy and Ali, 2011) and expression changes of SR genes in plants can affect the choice of splice sites, resulting in changes in splicing patterns (Lopato *et al.*, 1999; Kalyna *et al.*, 2003). We identified

20 SR proteins classified into six subfamilies plus three SR-like genes (Figure S8). The number of SR proteins in *N. attenuata* is greater than in *A. thaliana* (Cruz *et al.*, 2014), mainly due to an additional copy in each of the SR, SR-like and RS subfamilies (Figure S8, Table S4). Among all identified SR and SR-like genes, seven (*NaSR34a-2*, *NaRSZ21*, *NaSC35*, *NaSCL30a-1*, *NaSCL33*, *NaRS31-1* and *NaSR45a-1*) were significantly upregulated in *N. attenuata* roots but none in leaves after 5 h of feeding by *M. sexta* (Table S4), which is consistent with the observation that AS responses are more strongly induced in roots than in leaves. The most upregulated SR or SR-like gene is *NaSR45a-1* (Table S4), which increased more than four-fold in roots. Functional studies showed that Arabidopsis SR45a (*AtSR45a*) interacts directly with both splicing factors for the initial definition of 3′ (U1-70K) and 5′ (U2AF^{35}b) splice sites and also interacts with other SR genes such as *AtSCL28* and *AtSR45* (Tanabe *et al.*, 2009). In addition, three other SR genes (*NaSCL33*, *NaRSZ21* and *NaSC35*) also showed more than 1.8-fold upregulation. In *A. thaliana*, the orthologue of *NaSCL33* (*AtSCL33*) binds to a specific intron sequence motif and generates the intron splicing variants (Thomas *et al.*, 2012), and the expression and splicing were also regulated after virus infection in *Brachypodium distachyon* (Lopato *et al.*, 1999). Both orthologues of *NaRSZ21* (*AtRSZ21*) and *NaSC35-2* (*AtSC35-like*) interact with the splicing factor of 3′ splicing sites (U1-70K) (Wu and Maniatis, 1993; Golovkin and Reddy, 1998). In addition, after 9 h of attack by *M. sexta*, the expression of all four of these SR and SR-like genes (*NaSR45a-1*, *NaSCL33*, *NaRSZ21* and *NaSC35*) decreased in roots to the control levels at 5 h (Table S4), a result consistent with the similar overall PSI changes based on all existing AS events (Figure S7b). Due to the lack of control samples at 9 h, we cannot conclude that a longer duration of feeding actually suppresses SR gene expression; however, their strong correlation suggests a likely causal relationship between increased SR gene expression and enhanced AS responses in roots.

Changes in protein phosphorylation and splicing patterns in SR and SR-related genes are also known to regulate AS (Lopato *et al.*, 1999; Ali *et al.*, 2003; Kalyna *et al.*, 2003; Isshiki *et al.*, 2006). *Manduca sexta* attack didn't strongly affect either the expression or the splicing of SR genes in leaves (Tables S4 and S5), indicating that the *M. sexta*-induced changes in AS in leaves might be due to changes in phosphorylation of SR proteins. In addition, regulation of AS can also be regulated by small RNAs (smRNA) (Jones-Rhoades *et al.*, 2006). Simulated herbivore attack can induce changes in smRNA populations in *N. attenuata* leaves (Pandey *et al.*, 2008), which might also have contributed to the observed AS responses in leaves and roots. However, further experiments that directly manipulate the expressions of the SR genes and smRNA populations are required.

Jasmonate signaling is involved in upregulation of SR genes in roots induced by *M. sexta* attack

Stress-induced changes in gene expression in plants are often mediated by phytohormones (Erb *et al.*, 2012; Iqbal *et al.*, 2014). Using JA-deficient transgenic plants (ir*AOC*), this analysis showed that the JA signaling pathway is involved in the herbivory-induced upregulation of the four SR genes in *N. attenuata* roots (Figure 4), although the detailed mechanisms differed. *NaRSZ21* and *NaSR45a-1* were induced in the WT by *M. sexta* attack but not in the roots of JA-deficient plants, both with and without MeJA supplementation (Figure 4), indicating that they were likely directly regulated by the JA signaling pathway. Interestingly, the expression of *NaRSZ21* remained low in the roots of MeJA-supplemented ir*AOC* plants, indicating that other jasmonates, such as 12-oxo-phytodienoic acid (OPDA) might be involved in its upregulation.

In roots, *M. sexta* attack increased the expression of *NaSC35* and *NaSCL33* in JA-deficient (ir*AOC*) plants only when these plants were supplemented with MeJA (Figure 4d, e), indicating that JA is required for the herbivory-induced upregulation of *NaSC35* and *NaSCL33*. However, JA itself is not sufficient to induce the expression changes of these two genes, since their expression remained the same after MeJA supplementation in JA-deficient plants in the absence of herbivore attack (Figure 4). Thus, we proposed two possible non-exclusive mechanisms for the role of JA in herbivory-induced upregulation of *NaSC35* and *NaSCL33*: (i) herbivory induces JA-independent defense responses in the leaves, which together with JA and its derivatives activate some mobile signaling molecules in the leaves that are transported to the roots and then directly result in the increased expression of *NaSC35* and *NaSCL33*; (ii) the herbivory-induced mobile signaling molecules in the leaves are transported to the roots to interact with root JA and its derivatives to increase the expression of *NaSC35* and *NaSCL33*. The key to disentangling these two possible mechanisms is to identify whether the JA signaling in roots is required for the upregulation of these two SR genes experiments that can be conducted with chimeric plants consisting of empty vector (EV) shoots grafted with JA-deficient roots (ir*AOC*) (Fragoso *et al.*, 2011).

In summary, these results show that herbivory-induced JA signaling is involved in regulating the herbivory-induced upregulation of four SR genes in the roots that are likely associated with the AS responses in this systemic tissue.

CONCLUSION

In conclusion, this analysis has provided evidence that insect herbivory not only elicits changes in gene expression in plants, but also induces genome-wide gene splicing responses in both leaves (local tissue) and roots (systemic

tissue), with stronger effects in roots. The induced AS responses are precisely regulated and likely contribute to transcriptomic fine tuning and anti-herbivore defenses and/or tolerance. The strong AS responses elicited in *N. attenuata* roots when leaves were attacked by *M. sexta* were likely due to the JA-dependent upregulation of several SR and SR-like genes that may interact with other splicing factors. This analysis suggests that an above-ground biotic stress can induce strong AS responses in a below-ground tissue. Furthermore, this analysis provides a foundation for future studies to understand the molecular mechanisms and functions of stress-induced AS in plants.

EXPERIMENTAL PROCEDURES

Plant material

The seeds of 30th-generation of inbred WT) *N. attenuata* and the isogenic transgenic line impaired in JA biosynthesis (ir*AOC*, A-07-457-1; Kallenbach *et al.*, 2012) were germinated according to the previously established protocol (Krügel *et al.*, 2002). Ten-day-old seedlings were transferred to small pots (TEKU JP 3050 104 pots; Pöppelmann, https://www.poeppelmann.com/) with Klasmann plug soil (Klasmann-Deilmann, http://www.klasmann-deilmann.com/). After 10 days, seedlings were transplanted to 1-L pots with sand to facilitate root sampling. Plants were grown in the glasshouse with a day/night cycle of 16 h (26–28°C)/8 h (22–24°C) under supplemental light from Master Sun-T PIA Agro 400 or Master Sun-T PIA Plus 600-W sodium lights (Philips Son-T Agro, http://www.lighting.philips.com). For MeJA supplementation of ir*AOC* plants, 150 μg MeJA was dissolved in 20 μl of lanolin paste (Baldwin *et al.*, 1996; Pluskota *et al.*, 2007; Schafer *et al.*, 2015); using a small spatula this was then rubbed on the base of the +1 leaf node of each plant for 24 h. For the control group, 20 μl of lanolin paste without MeJA was used.

Manduca sexta feeding treatment, RNA isolation and sequencing

Eggs of the tobacco hornworm *M. sexta* were obtained from an in-house colony, in which insects were reared in a growth chamber (Snijders Scientific, http://www.snijders-tilburg.nl/) at 26°C/16-h light, 24°C/8-h dark, and 65% relative humidity. The freshly hatched larvae were first reared on fully developed rosette leaves of *N. attenuata* plants for 18 h. Then all larvae were starved for 1 h in a Petri dish before they were placed on the first fully elongated leaf (+1 position) of randomly selected rosette stage plants. Three larvae of similar size were placed on each plant at the same time (9 a.m.). The plants without larvae were considered as controls. The larvae were allowed to feed for 5 and 9 h, respectively. After treatment, both leaf and root samples were collected from plants with and without *M. sexta* feeding. Leaves were directly frozen in liquid nitrogen, whereas roots were first washed in a water tank for a few seconds to remove sand before they were frozen in liquid nitrogen. Three biological replicates were harvested for each treatment and control group.

Total RNA was extracted from leaves or roots tissues using TRIZOL (Invitrogen, http://www.invitrogen,com/) according to the manufacturer's protocol. All RNA samples were subsequently treated with RNase-free DNase-I (Fermentas, https://www.thermofisher.com/) to remove any contamination from genomic DNA. The mRNA was enriched using an mRNA-seq sample preparation kit (Illumina, https://www.illumina.com/), and insertion libraries of around 200 bp were constructed using an Illumina whole transcriptome analysis kit following the standard protocol (Illumina HiSeq system). All libraries were then sequenced on the Illumina HiSeq 2000 at Beijing Genomics Institute (BGI), Hong Kong. On average, 48 million 90-nucleotide paired-end raw reads for each sample were obtained (Table S1).

Read mapping, sequencing depth estimation, transcript assembly and abundance estimation, and identification of differentially expressed genes/transcripts

An overview of bioinformatics pipeline used for data analysis is given in Figure S12. The raw sequence reads were trimmed using ADAPTERREMOVAL (v.1.1) (Lindgreen, 2012) with the parameters '–collapse –trimns –trimqualities 2 –minlength 36'. The trimmed reads were then aligned to the *N. attenuata* genome assembly (v.1.0) using TOPHAT2 (v.2.0.6) (Trapnell *et al.*, 2009), with maximum and minimum intron size set to 50 000 and 41 bp, respectively, estimated from the *N. attenuata* genome annotation. The numbers of uniquely mapped reads and SJ mapped reads were then counted using SAMTOOLS (v.0.1.19) (Li *et al.*, 2009).

For each BAM file, JUNCBASE v.0.6 (Brooks *et al.*, 2011) was used to count the number of supporting reads for each detected junction. Only junctions mapped by more than 10 reads were used in the final analysis. The transcripts were assembled using CUFFLINKS (v.2.1.1) (Trapnell *et al.*, 2012) with the *N. attenuata* genome annotation (v.1.0) as the reference. The ORF of each transcript was analyzed using TransDecoder from TRINITY v.20131110 (Grabherr *et al.*, 2011).

To estimate the expression level of assembled transcripts, all trimmed reads were re-mapped to the assembled transcripts using RSEM (v.1.2.8) (Li and Dewey, 2011). Transcripts per million (TPM) was calculated for each transcript (Wagner *et al.*, 2012). Only those genes with TPM greater than five in at least three samples were considered for downstream comparative analysis at the gene level. Similarly, only those transcripts with TPM greater than one and comprising at least 10% of the transcripts in at least three samples were used in the comparative analysis at the transcript level.

The DE genes/transcripts between *M. sexta* attacked and control samples were identified using the EDGER package (v.3.8.4) (Robinson *et al.*, 2010) with the reads count matrix estimated from RSEM as described above. Genes/transcripts with more than two-fold changes and a false discovery rate (FDR)-adjusted *P*-value less than 0.01 were considered as DE genes/transcripts (Figure S12).

Detection of AS and *M. sexta* attack-induced DS genes

All AS analyses were based on SJs obtained from the BAM files produced by TOPHAT2, and the overview of tools used for AS analysis is depicted in Figure S12. To remove the false-positive junctions that were likely due to non-specific or erroneous alignments, all original junctions were filtered based on an overhang size greater than 13 bp, estimated based on an approach similar to Cui *et al.* (2014). In brief, we first constructed 89 484 sequences from the join each side of a known SJ (positive dataset) and joined each side of randomly assigned SJs from exons of different genes (negative dataset). For each sequence, 79 bp from either side of the SJ were used, which gives a 158-bp SJ sequence that ensured a 11-bp overhang, since the length of our sequencing reads is 90 bp. Then all reads were mapped to both positive and negative datasets using BWA (v.0.6.1) (Li and Durbin, 2009). The reads that aligned to the negative dataset were considered false positives. We evaluated two parameters, the overhang size and the number of supporting reads, to determine how these two factors affect the

false-positive rate. Overall, when an overhang size greater than 13 bp was required, the false-positive rate dropped to 0, while the true-positive rate remained at 98.7% (Figure S13). When the cutoff number of supporting reads was set to 10, the false-positive rate dropped to 5.9% (Figure S13); however, the true-positive rate also decreased (to 39.4%). To minimize the false-positive rate and maximize the true-positive rate, we used an overhang size greater than 13 bp as the filtering criterion.

All filtered SJs were then used for identification of AS events and annotation using JUNCBASE v.0.6 (Brooks *et al.*, 2011). The DS events were identified based on the PSI calculated using JUNCBASE v.0.6 (Brooks *et al.*, 2011) in each sample. The PSI value ranged from zero to one, which represents the relative ratio of isoforms generated from AS events. Student's *t*-test was used to compare the PSI values between *M. sexta* feeding and control samples to identify DS AS events, and the *P*-value cutoff was set as 0.05. To reduce false positives, we only considered those AS events with the biggest PSI difference (> 3%) between the samples from the two groups being compared. To select an AS event that showed no difference in total read coverage between control and *M. sexta* feeding treatment, the generalized linear model (GLM) was used with a *P*-value cutoff set as greater than 0.05. All filtering and statistical analyses were performed in R 3.0.2 (R Development Core Team 2013).

Identification of SR protein genes

The coding sequences (CDSs) of SR and SR-like genes in *A. thaliana* were obtained from the TAIR database (https://www.arabidopsis.org/index.jsp) based on the gene ID listed in Cruz *et al.* (2014). All sequences were used to find homologous genes in *N. attenuata* using BLAST (v.2.2.25) (Altschul *et al.*, 1990). A python script was used to filter the BLAST results based on the following requirements: identity > 60%, alignment coverage > 60% of the query length, E-value < 1×10^{-10}. The CDSs of *N. attenuata* and *A. thaliana* were aligned using MUSCLE (v.3.8.31) (Edgar, 2004) based on the protein sequences translated by TRANSLATORX (v.1.1) (Abascal *et al.*, 2010) and all non-informational sites (gaps in more than 20% of the sequences) were removed by TRIMAL (v.1.4) (Capella-Gutierrez *et al.*, 2009). The aligned sequences were then used to construct a phylogenetic tree with PHYML (v.20140206) (Guindon *et al.*, 2010), with a molecular evolution model being estimated by JMODELTEST2 (v.2.1.5). Accession numbers for the *Nicotiana attenuata* SR genes are shown in Table S6.

Quantitative RT-PCR

All cDNA samples were synthesized from 1 μg total RNA using SuperScript II reverse transcriptase (Thermo Fisher Scientific, https://www.thermofisher.com). The relative level of transcripts accumulated in selected AS events was measured using qPCR on a Stratagene MX3005P PCR cycler (Stratagene, http://www.stratagene.com/). For all qPCRs, the elongation factor-1A gene, *NaEF1a* (accession number D63396), was used as the internal standard for normalization as described in Oh *et al.* (2013). In order to measure the expression of detected transcript variants within the same gene, we designed transcript-specific primers (listed in Table S7). All qPCR reactions were performed using a qPCR core kit for SYBR Green I (Eurogentec, http://www.eurogentec.com/) in a 20-μl reaction system. Three biological replicates were used for all qPCR measurements.

Gene Ontology and transcript domain analysis

The GO annotations of all *N. attenuata* genes were obtained from *N. attenuata* genome annotation. The functional enrichment analysis was performed using CLUGO v.2.1.1 (Bindea *et al.*, 2009). The over-represented GO terms were computed by comparing the GO of the custom gene set with the reference GO annotation set, and the GO fusion function term was used. The significance levels of the enriched GO terms were determined by Bonferroni corrected *P*-values ($P < 0.01$). The protein domains for all genes were predicted by searching against the NCBI Conserved Domains database (CDD v.3.12) with an E-value < 0.01.

ACKNOWLEDGEMENTS

We thank T. Krügel and the greenhouse team for taking care of plants, T. Brockmüller and A. Navarro-Quezada for constructive discussions on data analysis, and P. Schlüter, E. Gaquerel, J. Gulati, M. Huber and J. Boyer for their constructive comments on the manuscript. We also thank E. Moulton for helping with sample collection and K. Gase for helping with data deposition. The work was supported by Swiss National Science Foundation (project number PEBZP3-142886 to SX), a Marie Curie Intra-European Fellowship (IEF) (project number 328935 to SX), Max Planck Society and European Research Council advanced grant ClockworkGreen (project number 293926 to ITB). The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Features of assembled transcripts.

Figure S2. Evaluations of sequencing depth.

Figure S3. The distribution of different types of alternative splicing events in *Nicotiana attenuata* leaves and roots.

Figure S4. Number of alternative splicing events that were unique or shared among different biological replicates in leaves and roots.

Figure S5. The distribution of reads coverage in the subset of splice junctions that show same expression between control and *Manduca sexta* feeding-induced samples.

Figure S6. The abundance of different types of alternative splicing events in the subset of splice junctions that were not differentially expressed after *Manduca sexta* attack.

Figure S7. Principal components analysis based on the percentage splicing index.

Figure S8. Phylogenetic tree of *Nicotiana attenuata* and Arabidopsis serine/arginine-rich (SR) and SR-like genes estimated using the maximum likelihood method.

Figure S9. Validation of expression of five selected serine/arginine-rich (SR) genes in leaves and roots using quantitative PCR.

Figure S10. Expression changes in *Manduca sexta* attack-induced transcripts and AS responses in the jasmonate (JA) signal pathway in roots.

Figure S11. The enriched Gene Ontology terms for up- and down-regulated genes in *Nicotiana attenuata* leaves and roots after 5 h of feeding by *Manduca sexta*.

Figure S12. Flowchart of the data analysis process.

Figure S13. Overhang size and reads coverage distributions in positive and negative splice junction datasets.

Table S1. Mapping results of RNA-sequencing data.

Table S2. Total number of annotated alternative splicing (AS) events and number of genes contained in these AS events.

Table S3. The expression of dominant and minor transcripts in differentially spliced (DS) genes induced by 5 h *M. sexta* attack and measured by quantitative PCR.

Table S4. The expression of all 23 serine/arginine-rich (SR) and SR-like genes in leaves and roots.

Table S5. Tables listing the percentage splicing index of all alternative splicing events identified in 23 serine/arginine-rich (SR) and SR-like genes in leaves and roots.

Table S6. The accession number of genes that were selected for alternative splicing and gene expression analysis.

Table S7. The primers used for validating differentially expressed serine/arginine-rich genes and differentially spliced transcripts.

Data S1. Tables listing all condition-specific alternative splicing events identified in leaves and roots.

Data S2. Detailed information of all identified differentially spliced alternative splicing genes in *Nicotiana attenuata* leaves and roots

REFERENCES

Abascal, F., Zardoya, R. and Telford, M.J. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, **38**, W7–W13.

Alexandrov, N.N., Troukhan, M.E., Brover, V.V., Tatarinova, T., Flavell, R.B. and Feldmann, K.A. (2006) Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. *Plant Mol. Biol.*, **60**, 69–85.

Ali, G.S., Golovkin, M. and Reddy, A.S. (2003) Nuclear localization and *in vivo* dynamics of a plant-specific serine/arginine-rich protein. *Plant J.* **36**, 883–893.

Ali, G.S., Palusa, S.G., Golovkin, M., Prasad, J., Manley, J.L. and Reddy, A.S. (2007) Regulation of plant developmental processes by a novel splicing factor. *PLoS ONE*, **2**, e471.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Baldwin, I.T., Schmelz, E.A. and Zhang, Z.P. (1996) Effects of octadecanoid metabolites and inhibitors on induced nicotine accumulation in *Nicotiana sylvestris*. *J. Chem. Ecol.*, **22**, 61–74.

Ballare, C.L. (2011) Jasmonate-induced defenses: a tale of intelligence, collaborators and rascals. *Trends Plant Sci.*, **16**, 249–257.

Barta, A., Kalyna, M. and Reddy, A.S.N. (2010) Implementing a rational and consistent nomenclature for serine/arginine-rich protein splicing factors (SR proteins) in plants. *Plant Cell*, **22**, 2926–2929.

Bilgin, D.D., Zavala, J.A., Zhu, J., Clough, S.J., Ort, D.R. and DeLucia, E.H. (2010) Biotic stress globally downregulates photosynthesis genes. *Plant, Cell Environ.*, **33**, 1597–1613.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pages, F., Trajanoski, Z. and Galon, J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.

Brooks, A.N., Yang, L., Duff, M.O., Hansen, K.D., Park, J.W., Dudoit, S., Brenner, S.E. and Graveley, B.R. (2011) Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res.*, **21**, 193–202.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. and Buell, C.R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genom.*, **7**, 327.

Capella-Gutierrez, S., Silla-Martinez, J.M. and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Chang, Y.F., Imam, J.S. and Wilkinson, M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.*, **76**, 51–74.

Chung, H.S., Koo, A.J., Gao, X., Jayanty, S., Thines, B., Jones, A.D. and Howe, G.A. (2008) Regulation and Function of *Arabidopsis JASMONATE ZIM*-Domain Genes in Response to Wounding and Herbivory. *Plant Physiol.*, **146**, 952–964.

Cruz, T.M., Carvalho, R.F., Richardson, D.N. and Duque, P. (2014) Abscisic acid (ABA) regulation of *Arabidopsis* SR protein gene expression. *int. J. Mol. Sci.*, **15**, 17541–17564.

Cui, P., Zhang, S., Ding, F., Ali, S. and Xiong, L. (2014) Dynamic regulation of genome-wide pre-mRNA splicing and stress tolerance by the Sm-like protein LSm5 in *Arabidopsis*. *Genome Biol.*, **15**, R1.

Dinesh-Kumar, S.P. and Baker, B.J. (2000) Alternatively spliced *N* resistance gene transcripts: their possible role in tobacco mosaic virus resistance. *Proc. Natl Acad. Sci. USA*, **97**, 1908–1913.

Ding, F., Cui, P., Wang, Z., Zhang, S., Ali, S. and Xiong, L. (2014) Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in *Arabidopsis*. *BMC Genom.*, **15**, 431.

Dinh, S.T., Baldwin, I.T. and Galis, I. (2013) The *HERBIVORE ELICITOR-REGULATED1* gene enhances abscisic acid levels and defenses against herbivores in *Nicotiana attenuata* plants. *Plant Physiol.*, **162**, 2106–2124.

Duque, P. (2011) A role for SR proteins in plant stress responses. *Plant Signal. Behav.*, **6**, 49–54.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Erb, M., Lenk, C., Degenhardt, J. and Turlings, T.C. (2009) The underestimated role of roots in defense against leaf attackers. *Trends Plant Sci.*, **14**, 653–659.

Erb, M., Meldau, S. and Howe, G.A. (2012) Role of phytohormones in insect-specific plant reactions. *Trends Plant Sci.*, **17**, 250–259.

Farmer, E.E. and Ryan, C.A. (1990) Interplant communication - airborne methyl jasmonate induces synthesis of proteinase-inhibitors in plant-leaves. *Proc. Natl Acad. Sci. USA*, **87**, 7713–7716.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K. and Mockler, T.C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.*, **20**, 45–58.

Fragoso, V., Goddard, H., Baldwin, I.T. and Kim, S.G. (2011) A simple and efficient micrografting method for stably transformed *Nicotiana attenuata* plants to examine shoot-root signaling. *Plant Methods*, **7**, 34.

Fragoso, V., Rothe, E., Baldwin, I.T. and Kim, S.G. (2014) Root jasmonic acid synthesis and perception regulate folivore-induced shoot metabolites and increase *Nicotiana attenuata* resistance. *New Phytol.*, **202**, 1335–1345.

de la Fuente van Bentem, S., Anrather, D., Roitinger, E., Djamei, A., Hufnagl, T., Barta, A., Csaszar, E., Dohnal, I., Lecourieux, D. and Hirt, H. (2006) Phosphoproteomics reveals extensive *in vivo* phosphorylation of *Arabidopsis* proteins involved in RNA metabolism. *Nucleic Acids Res.*, **34**, 3267–3278.

Fulton, D.C., Stettler, M., Mettler, T. *et al.* (2008) β-AMYLASE4, a Noncatalytic Protein Required for Starch Breakdown, Acts Upstream of Three Active β-Amylases in *Arabidopsis* Chloroplasts. *Plant Cell*, **20**, 1040–1058.

Gilardoni, P.A., Schuck, S., Jungling, R., Rotter, B., Baldwin, I.T. and Bonaventure, G. (2010) SuperSAGE analysis of the *Nicotiana attenuata* transcriptome after fatty acid-amino acid elicitation (FAC): identification of early mediators of insect responses. *BMC Plant Biol.*, **10**, 66.

Golovkin, M. and Reddy, A.S.N. (1998) The plant U1 small nuclear ribonucleoprotein particle 70K protein interacts with two novel serine/arginine-rich proteins. *Plant Cell*, **10**, 1637–1647.

Golovkin, M. and Reddy, A.S. (1999) An SC35-like protein and a novel serine/arginine-rich protein interact with *Arabidopsis* U1-70K protein. *J. Biol. Chem.*, **274**, 36428–36438.

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

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.

Gulati, J., Kim, S.G., Baldwin, I.T. and Gaquerel, E. (2013) Deciphering herbivory-induced gene-to-metabolite dynamics in *Nicotiana attenuata* tissues using a multifactorial approach. *Plant Physiol.*, **162**, 1042–1059.

Gulati, J., Baldwin, I.T. and Gaquerel, E. (2014) The roots of plant defenses: integrative multivariate analyses uncover dynamic behaviors of gene and metabolic networks of roots elicited by leaf herbivory. *Plant J.* **77**, 880–892.

Howard, B.E., Hu, Q.W., Babaoglu, A.C. *et al.* (2013) High-throughput RNA sequencing of *Pseudomonas*-infected *Arabidopsis* reveals hidden transcriptome complexity and novel splice variants. *PLoS ONE*, **8**, e74183.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A. and Shinozaki, K. (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res.*, **32**, 5096–5103.

Iqbal, N., Umar, S., Khan, N.A. and Khan, M.I.R. (2014) A new perspective of phytohormones in salinity tolerance: Regulation of proline metabolism. *Environ. Exp. Bot.*, **100**, 34–42.

Isshiki, M., Tsumoto, A. and Shimamoto, K. (2006) The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. *Plant Cell*, **18**, 146–158.

James, A.B., Syed, N.H., Bordage, S., Marshall, J., Nimmo, G.A., Jenkins, G.I., Herzyk, P., Brown, J.W. and Nimmo, H.G. (2012) Alternative splicing mediates responses of the *Arabidopsis* circadian clock to temperature changes. *Plant Cell*, **24**, 961–981.

Jones-Rhoades, M.W., Bartel, D.P. and Bartel, B. (2006) MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, **57**, 19–53.

Jurica, M.S. and Moore, M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell*, **12**, 5–14.

Kallenbach, M., Bonaventure, G., Gilardoni, P.A., Wissgott, A. and Baldwin, I.T. (2012) Empoasca leafhoppers attack wild tobacco plants in a jasmonate-dependent manner and identify jasmonate mutants in natural populations. *Proc. Natl Acad. Sci. USA*, **109**, E1548–E1557.

Kalyna, M., Lopato, S. and Barta, A. (2003) Ectopic Expression of atRSZ33 Reveals Its Function in Splicing and Causes Pleiotropic Changes in Development. *Mol. Biol. Cell*, **14**, 3565–3577.

Kandoth, P.K., Ranf, S., Pancholi, S.S., Jayanty, S., Walla, M.D., Miller, W., Howe, G.A., Lincoln, D.E. and Stratmann, J.W. (2007) Tomato MAPKs LeMPK1, LeMPK2, and LeMPK3 function in the systemin-mediated defense response against herbivorous insects. *Proc. Natl Acad. Sci. USA*, **104**, 12205–12210.

Kazan, K. and Manners, J.M. (2008) Jasmonate signaling: toward an integrated view. *Plant Physiol.*, **146**, 1459–1468.

Kesari, R., Lasky, J.R., Villamor, J.G., Marais, D.L.D., Chen, Y.J.C., Liu, T.W., Lin, W., Juenger, T.E. and Verslues, P.E. (2012) Intron-mediated alternative splicing of *Arabidopsis P5CS1* and its association with natural variation in proline and climate adaptation. *Proc. Natl Acad. Sci. USA*, **109**, 9197–9202.

Kessler, A., Halitschke, R. and Baldwin, I.T. (2004) Silencing the jasmonate cascade: induced plant defenses and insect populations. *Science*, **305**, 665–668.

Kim, S.G., Yon, F., Gaquerel, E., Gulati, J. and Baldwin, I.T. (2011) Tissue specific diurnal rhythms of metabolites and their regulation during herbivore attack in a native tobacco, *Nicotiana attenuata*. *PLoS One*, **6**, e26214.

Krügel, T., Lim, M., Gase, K., Halitschke, R. and Baldwin, I.T. (2002) *Agrobacterium*-mediated transformation of *Nicotiana attenuata*, a model ecological expression system. *Chemoecology*, **12**, 177–183.

Lazar, G., Schaal, T., Maniatis, T. and Goodman, H.M. (1995) Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor Sf2/Asf. *Proc. Natl Acad. Sci. USA*, **92**, 7672–7676.

Leviatan, N., Alkan, N., Leshkowitz, D. and Fluhr, R. (2013) Genome-wide survey of cold stress regulated alternative splicing in *Arabidopsis thaliana* with tiling microarray. *PLoS ONE*, **8**, e66511.

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, W., Lin, W.D., Ray, P., Lan, P. and Schmidt, W. (2013) Genome-wide detection of condition-sensitive alternative splicing in *Arabidopsis* roots. *Plant Physiol.*, **162**, 1750–1763.

Lin, W.Y., Matsuoka, D., Sasayama, D. and Nanmori, T. (2010) A splice variant of *Arabidopsis* mitogen-activated protein kinase and its regulatory function in the MKK6-MPK13 pathway. *Plant Sci.*, **178**, 245–250.

Lindgreen, S. (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes*, **5**, 337.

Lopato, S., Kalyna, M., Dorner, S., Kobayashi, R., Krainer, A.R. and Barta, A. (1999) atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes. *Genes Dev.*, **13**, 987–1001.

Machado, R.A., Ferrieri, A.P., Robert, C.A., Glauser, G., Kallenbach, M., Baldwin, I.T. and Erb, M. (2013) Leaf-herbivore attack reduces carbon reserves and regrowth from the roots via jasmonate and auxin signaling. *New Phytol.*, **200**, 1234–1246.

Manley, J.L. and Krainer, A.R. (2010) A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes Dev.*, **24**, 1073–1074.

Maron, J.L. and Crone, E. (2006) Herbivory: effects on plant abundance, distribution and population growth. *Proc. Biol. Sci.*, **273**, 2575–2584.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.*, **22**, 1184–1195.

Mastrangelo, A.M., Marone, D., Laido, G., De Leonardis, A.M. and De Vita, P. (2012) Alternative splicing: enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci.*, **185**, 40–49.

Melamud, E. and Moult, J. (2009) Stochastic noise in splicing machinery. *Nucleic Acids Res.*, **37**, 4873–4886.

Meldau, S., Wu, J. and Baldwin, I.T. (2009) Silencing two herbivory-activated MAP kinases, SIPK and WIPK, does not increase *Nicotiana attenuata*'s susceptibility to herbivores in the glasshouse and in nature. *New Phytol.*, **181**, 161–173.

Mitra, S. and Baldwin, I.T. (2008) Independently silencing two photosynthetic proteins in *Nicotiana attenuata* has different effects on herbivore resistance. *Plant Physiol.*, **148**, 1128–1138.

Mitra, S. and Baldwin, I.T. (2014) RuBPCase activase (RCA) mediates growth-defense trade-offs: silencing RCA redirects jasmonic acid (JA) flux from JA-isoleucine to methyl jasmonate (MeJA) to attenuate induced defense responses in *Nicotiana attenuata*. *New Phytol.*, **201**, 1385–1395.

Oh, Y., Baldwin, I.T. and Galis, I. (2013) A jasmonate ZIM-domain protein NaJAZd regulates floral jasmonic acid levels and counteracts flower abscission in Nicotiana attenuata plants. *PLoS ONE*, **8**, e57868.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

Pandey, S.P., Shahi, P., Gase, K. and Baldwin, I.T. (2008) Herbivory-induced changes in the small-RNA transcriptome and phytohormone signaling in *Nicotiana attenuata*. *Proc. Natl Acad. Sci. USA*, **105**, 4559–4564.

Pluskota, W.E., Qu, N., Maitrejean, M., Boland, W. and Baldwin, I.T. (2007) Jasmonates and its mimics differentially elicit systemic defence responses in Nicotiana attenuata. *J. Exp. Bot.*, **58**, 4071–4082.

R Development Core Team (2013) *R: a language and environment for statistical computing*. Vienna, Austria: R Development Core Team.

Reddy, A.S.N. (2004) Plant serine/arginine-rich proteins and their role in pre-mRNA splicing. *Trends Plant Sci.*, **9**, 541–547.

Reddy, A.S.N. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu. Rev. Plant Biol.*, **58**, 267–294.

Reddy, A.S.N. and Ali, G.S. (2011) Plant serine/arginine-rich proteins: roles in precursor messenger RNA splicing, plant development, and stress responses. *Wiley Interdiscip. Rev. RNA*, **2**, 875–889.

Reymond, P., Bodenhausen, N., Van Poecke, R.M., Krishnamurthy, V., Dicke, M. and Farmer, E.E. (2004) A conserved transcript pattern in response to a specialist and a generalist herbivore. *Plant Cell*, **16**, 3132–3147.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Schafer, M., Meza-Canales, I.D., Navarro-Quezada, A., Brutting, C., Vankova, R., Baldwin, I.T. and Meldau, S. (2015) Cytokinin levels and signaling respond to wounding and the perception of herbivore elicitors in *Nicotiana attenuata*. *J. Integr. Plant Biol.*, **57**, 198–212.

Schwachtje, J. and Baldwin, I.T. (2008) Why does herbivore attack reconfigure primary metabolism? *Plant Physiol.*, **146**, 845–851.

Schwachtje, J., Minchin, P.E., Jahnke, S., van Dongen, J.T., Schittko, U. and Baldwin, I.T. (2006) SNF1-related kinases allow plants to tolerate herbivory by allocating carbon to roots. *Proc. Natl Acad. Sci. USA*, **103**, 12935–12940.

Shen, Y., Zhou, Z., Wang, Z. *et al.* (2014) Global Dissection of Alternative Splicing in Paleopolyploid Soybean. *Plant Cell*, **26**, 996–1008.

Steppuhn, A. and Baldwin, I.T. (2007) Resistance management in a native plant: nicotine prevents herbivores from compensating for plant protease inhibitors. *Ecol. Lett.*, **10**, 499–511.

Steppuhn, A., Gase, K., Krock, B., Halitschke, R. and Baldwin, I.T. (2004) Nicotine's defensive function in nature. *PLoS Biol.*, **2**, E217.

Tanabe, N., Yoshimura, K., Kimura, A., Yabuta, Y. and Shigeoka, S. (2007) Differential Expression of Alternatively Spliced mRNAs of *Arabidopsis* SR Protein Homologs, atSR30 and atSR45a, in Response to Environmental Stress. *Plant Cell Physiol.*, **48**, 1036–1049.

Tanabe, N., Kimura, A., Yoshimura, K. and Shigeoka, S. (2009) Plant-specific SR-related protein atSR45a interacts with spliceosomal proteins in plant nucleus. *Plant Mol. Biol.*, **70**, 241–252.

Thatcher, S.R., Zhou, W., Leonard, A., Wang, B.B., Beatty, M., Zastrow-Hayes, G., Zhao, X., Baumgarten, A. and Li, B. (2014) Genome-wide analysis of alternative splicing in *Zea mays*: landscape and genetic regulation. *Plant Cell*, **26**, 3472–3487.

Thomas, J., Palusa, S.G., Prasad, K.V.S.K., Ali, G.S., Surabhi, G.K., Ben-Hur, A., Abdel-Ghany, S.E. and Reddy, A.S.N. (2012) Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of *SCL33* pre-mRNA. *Plant J.* **72**, 935–946.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.

Vriet, C., Smith, A.M. and Wang, T.L. (2014) Root starch reserves are necessary for vigorous re-growth following cutting back in *Lotus japonicus*. *PLoS ONE*, **9**, e87333.

Wagner, G.P., Kin, K. and Lynch, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, **131**, 281–285.

Wang, B.B. and Brendel, V. (2004) The ASRG database: identification and survey of *Arabidopsis thaliana* genes involved in pre-mRNA splicing. *Genome Biol.*, **5**, R102.

Weaver, L.M., Swiderski, M.R., Li, Y. and Jones, J.D.G. (2006) The *Arabidopsis thaliana* TIR-NB-LRR R-protein, RPP1A; protein localization and constitutive activation of defence by truncated alleles in tobacco and *Arabidopsis*. *Plant J.* **47**, 829–840.

Will, C.L. and Luhrmann, R. (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.*, **3**, a003707.

Woldemariam, M.G., Onkokesung, N., Baldwin, I.T. and Galis, I. (2012) Jasmonoyl-l-isoleucine hydrolase 1 (JIH1) regulates jasmonoyl-l-isoleucine levels and attenuates plant defenses against herbivores. *Plant J.* **72**, 758–767.

Wu, J. and Baldwin, I.T. (2010) New insights into plant responses to the attack from insect herbivores. *Annu. Rev. Genet.*, **44**, 1–24.

Wu, J.Y. and Maniatis, T. (1993) Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, **75**, 1061–1070.

Wu, J., Hettenhausen, C., Meldau, S. and Baldwin, I.T. (2007) Herbivory rapidly activates MAPK signaling in attacked and unattacked leaf regions but not between leaves of *Nicotiana attenuata*. *Plant Cell*, **19**, 1096–1122.

Wu, J., Wang, L. and Baldwin, I.T. (2008) Methyl jasmonate-elicited herbivore resistance: does MeJA function as a signal without being hydrolyzed to JA? *Planta*, **227**, 1161–1168.

Wu, H.P., Su, Y.S., Chen, H.C., Chen, Y.R., Wu, C.C., Lin, W.D. and Tu, S.L. (2014) Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in *Physcomitrella patens*. *Genome Biol.*, **15**, R10.

Yan, L., Zhai, Q., Wei, J. *et al.* (2013) Role of tomato lipoxygenase D in wound-induced jasmonate biosynthesis and plant immunity to insect herbivores. *PLoS Genet.*, **9**, e1003964.

Zhang, Z.-P. and Baldwin, I.T. (1997) Transport of [2-^{14}C]jasmonic acid from leaves to roots mimics wound-induced changes in endogenous jasmonic acid pools in *Nicotiana sylvestris*. *Planta*, **203**, 436–441.

Zhang, Z.G., Xin, D.D., Wang, P., Zhou, L., Hu, L.D., Kong, X.Y. and Hurst, L.D. (2009) Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol.*, **7**, 23.

Zhang, G., Guo, G., Hu, X. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.*, **20**, 646–654.

Zhang, X., Rosen, B.D., Tang, H., Krishnakumar, V. and Town, C.D. (2015) Polyribosomal RNA-Seq reveals the decreased complexity and diversity of the *Arabidopsis* translatome. *PLoS ONE*, **10**, e0117699.

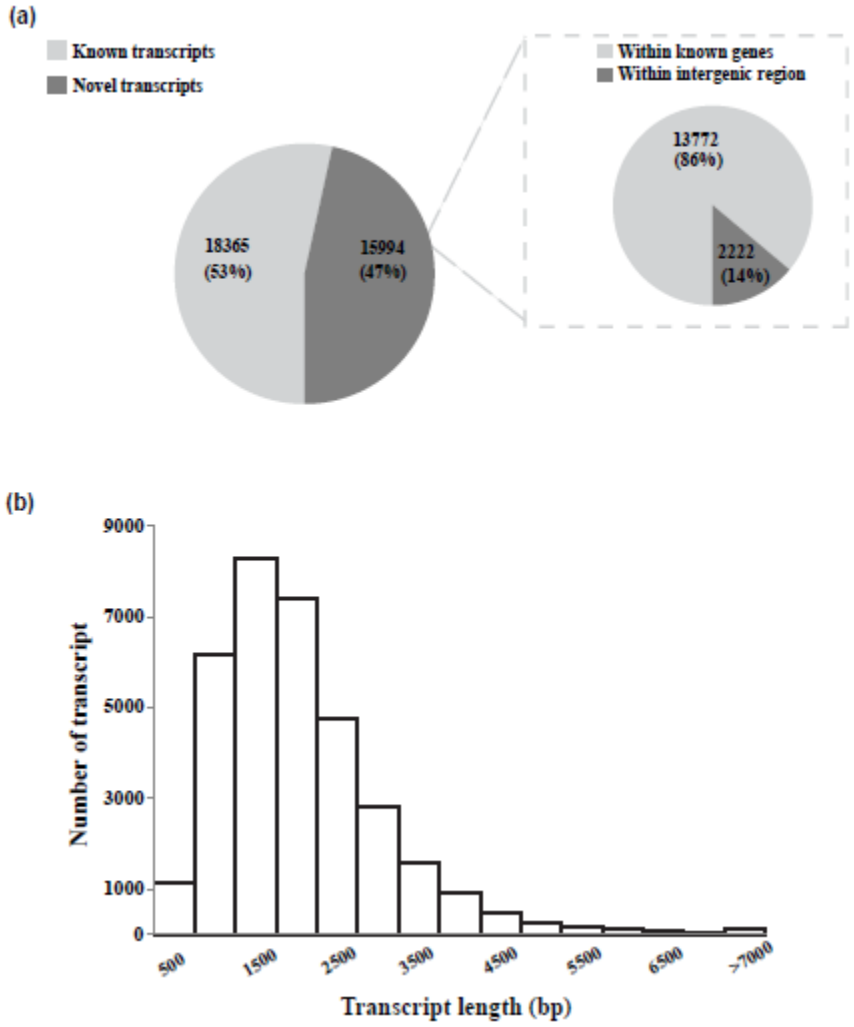


Figure S1. Features of assembled transcripts. (**a**), the distribution of assembled known and novel transcripts. The insert indicates the proportion of novel transcripts located within the genes region or inter-genetic region; (**b**), the length distribution of assembled transcripts.

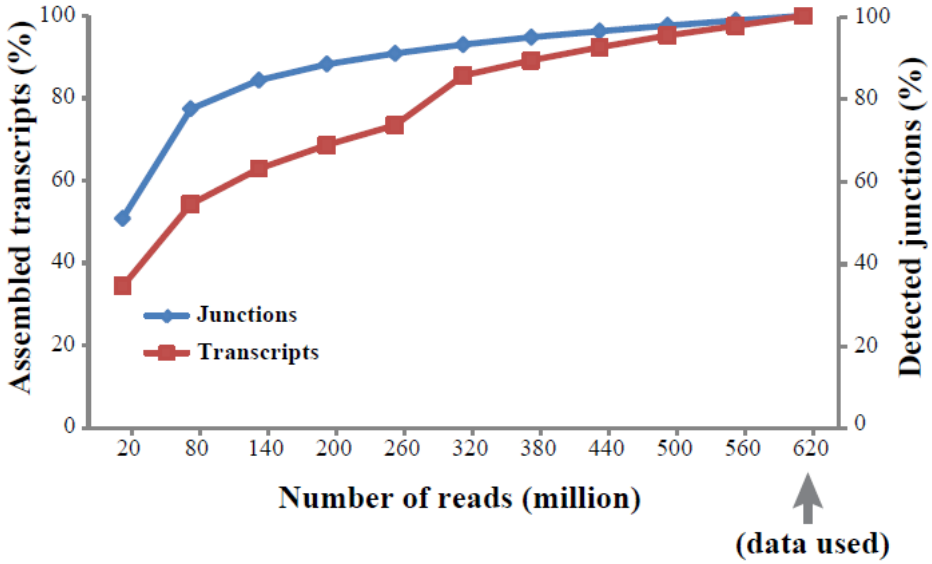


Figure S2. Evaluation on sequencing depth. The blue line represents the percentage of total detected junctions that were supported by more than 10 reads and the red line represents the percentage of total assembled novel transcripts. X-axis represents the number of sub-sampled reads that were used for analysis, and Y-axis refers to percentage of total assembled transcripts (left) and detected junctions (right). The gray arrow indicates the sequencing depth of data that were used for downstream AS analysis.

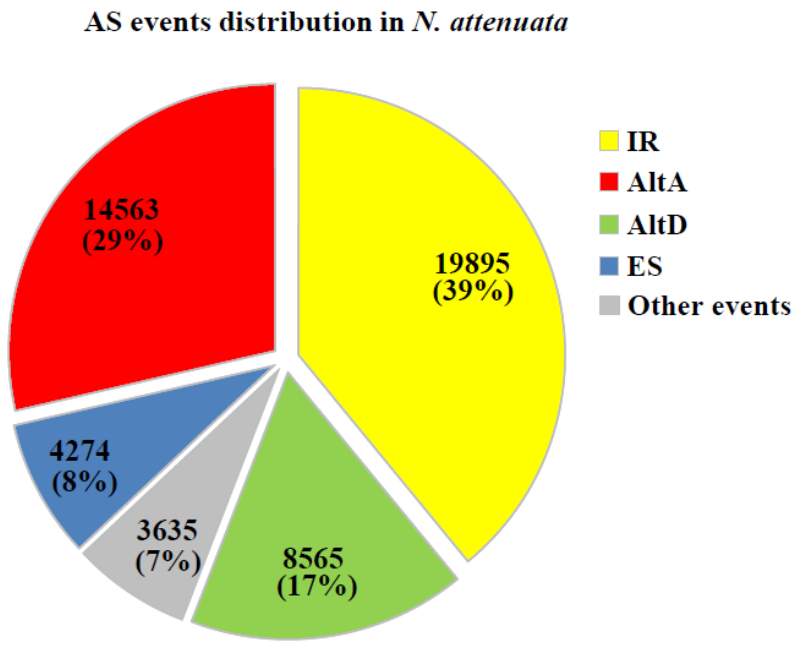


Figure S3. The distribution of different types of AS events in *N. attenuata* leaves and roots. IR: intron retention; AltA: 3' accepter site (AltA); AltD: alternative 5' donor site; ES: exon skipping. Other events: other types of AS that resulted from the combination of different above mentioned four types.

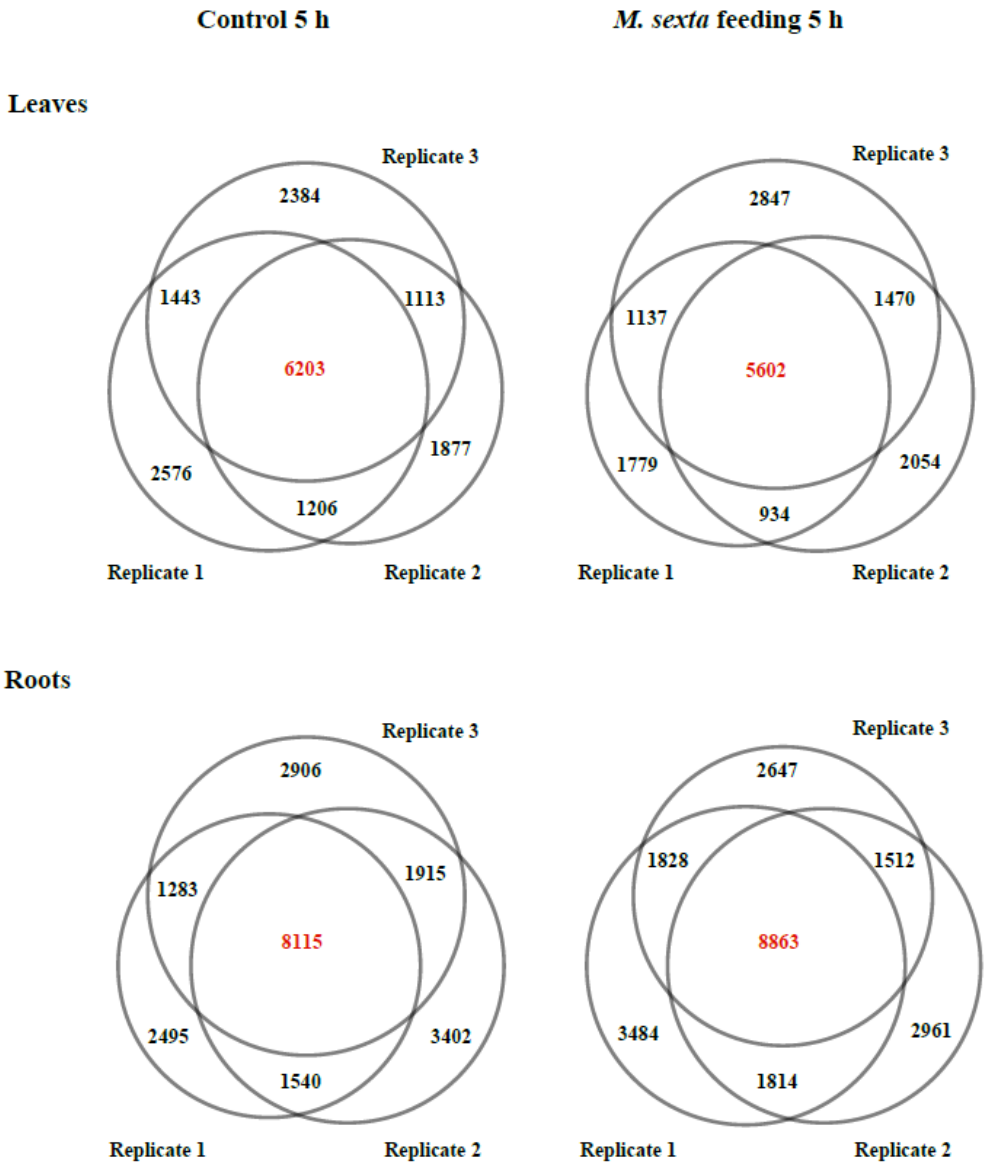


Figure S4. Number of AS events that were unique or shared among different biological replicates in leaves and roots. Each circle refers to one biological replicate, and numbers inside the circles refer to number of AS events. Top and bottom rows are leaves and roots samples, respectively; left and right columns are the control and *M. sexta* feeding samples at 5h, respectively.

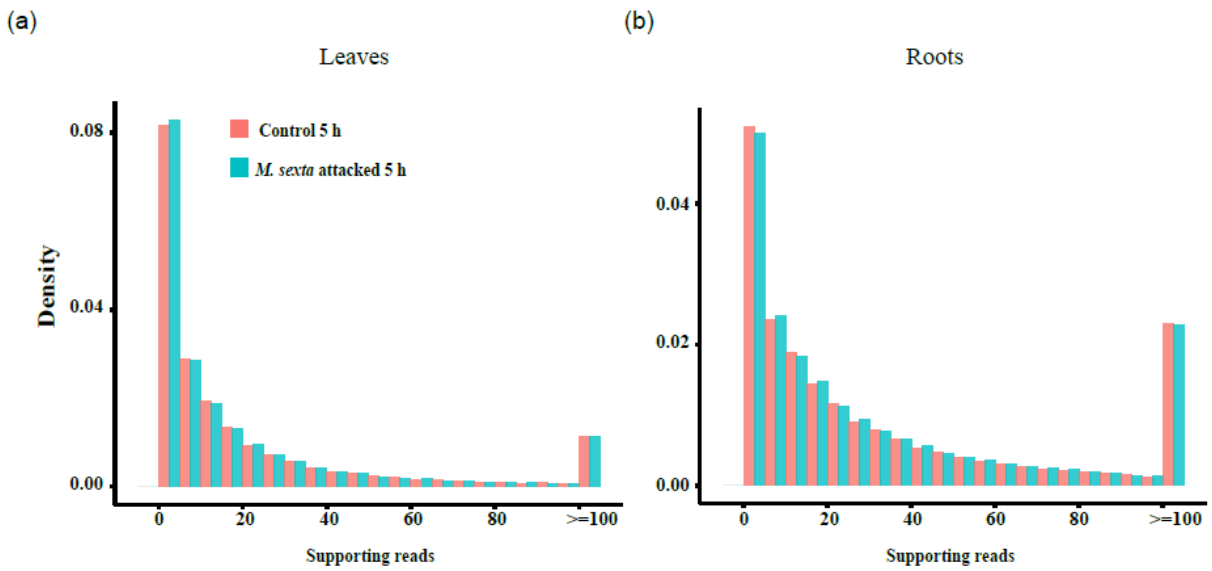


Figure S5. The distribution of reads coverage in the subset SJs that show same expression between control and *M. sexta* induced samples. The distribution of number of supporting reads to the junction area in leaves **(a)** and roots **(b)**. Y-axis refers to the density and X-axis refers to number of supporting reads. Red and green bars refer to control and *M. sexta* attacked samples at 5 h, respectively.

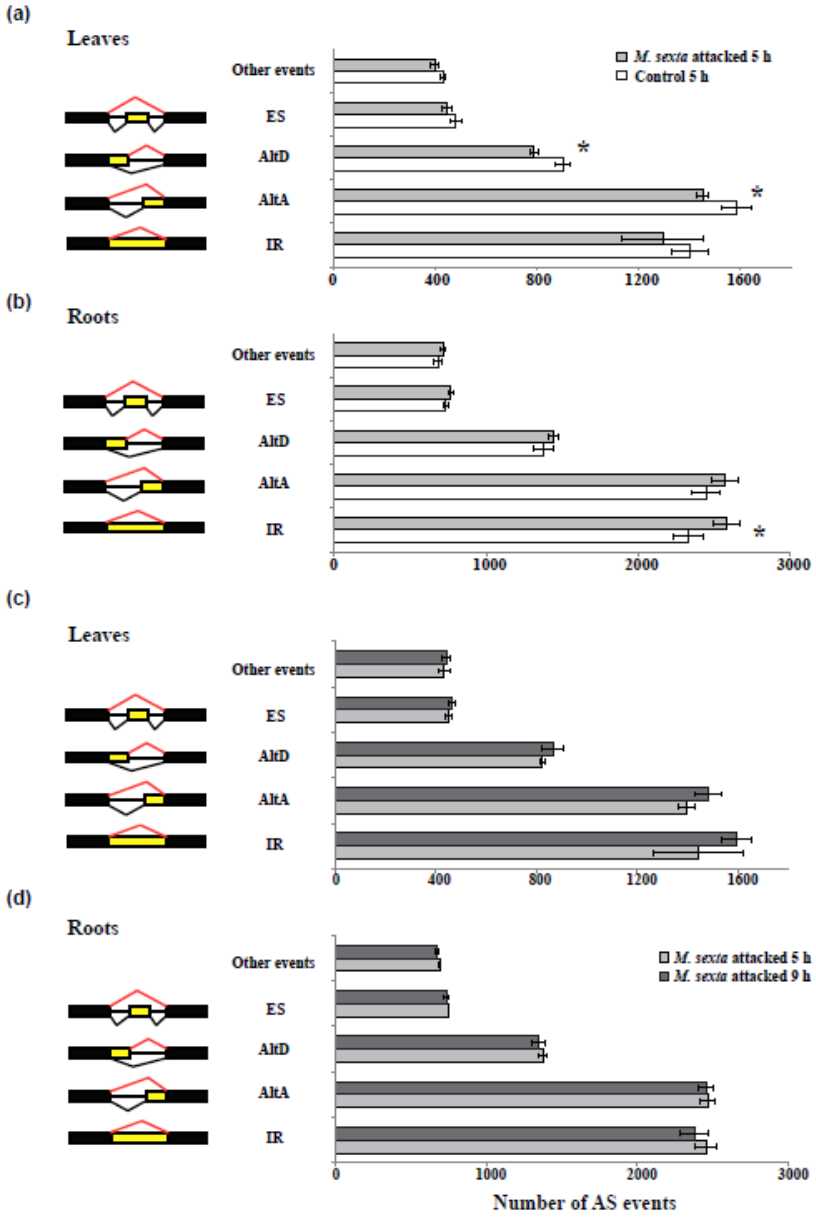


Figure S6. The abundance of different types of AS events in the subset of SJs that were not differentially expressed after *M. sexta* attack. (a) and **(b)**, the abundance of different types of AS events differing between control and *M. sexta* feeding 5 h samples in leaves (a) and roots (b); **(c) and (d)**, the abundance of different types of AS events in leaves (c) and roots (d) of plants with 5 and 9 h of *M. sexta* feeding. IR: intron retention; AltA: 3' accepter site (AltA); AltD: alternative 5' donor site; ES: exon skipping. The bar refers to number of AS events for each type. Asterisk indicates the significance as determined by Student's-*t* tests ($P < 0.05$). Error bars refer to standard error (SE).

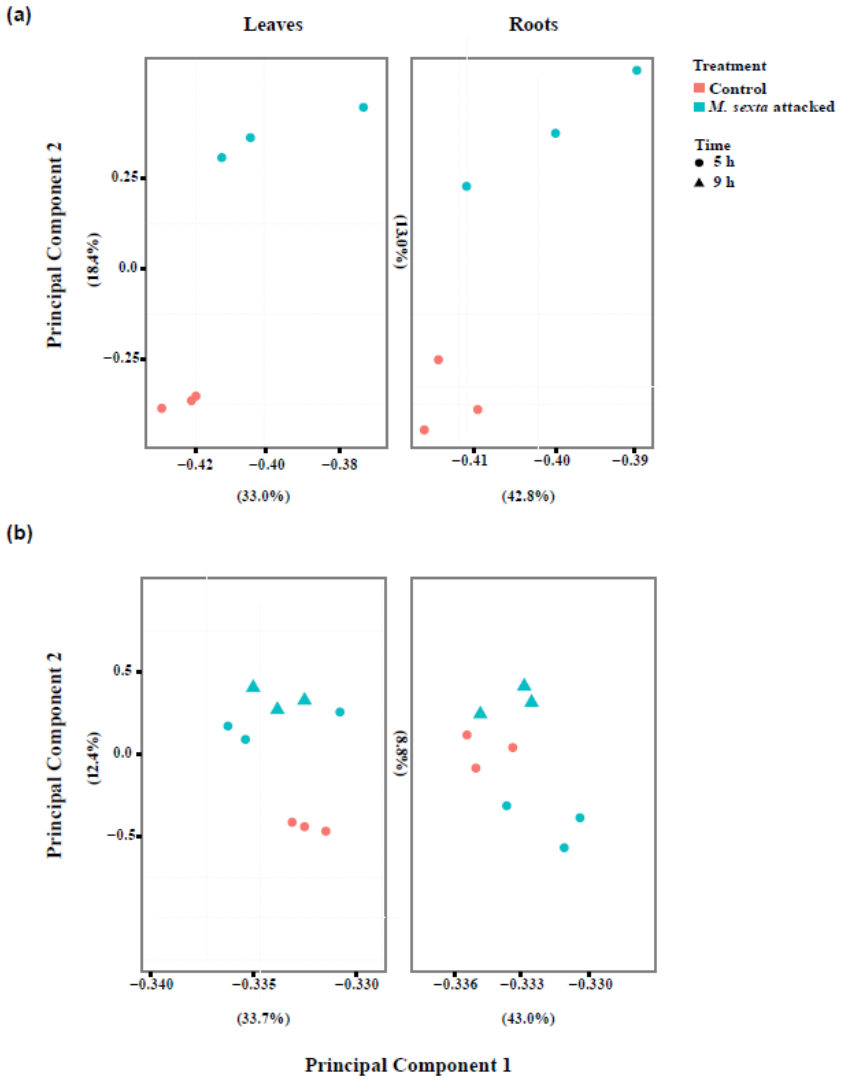


Figure S7. Principal component analysis (PCA) based on the percentage splicing index (PSI). (a), the global AS responses in leaves and roots after 5 h of *M. sexta* feeding; **(b)**, the global AS responses after 5 and 9 h of *M. sexta* feeding. The percentage of variance explained by each axis is shown in parentheses. Different colours refer to different treatments and different shapes refer to different duration of *M. sexta* feeding. Light blue refers to *M. sexta* feeding, and vermillion refers to control. Filled circle refer to samples harvested after 5 h and triangle refers to 9 h samples.

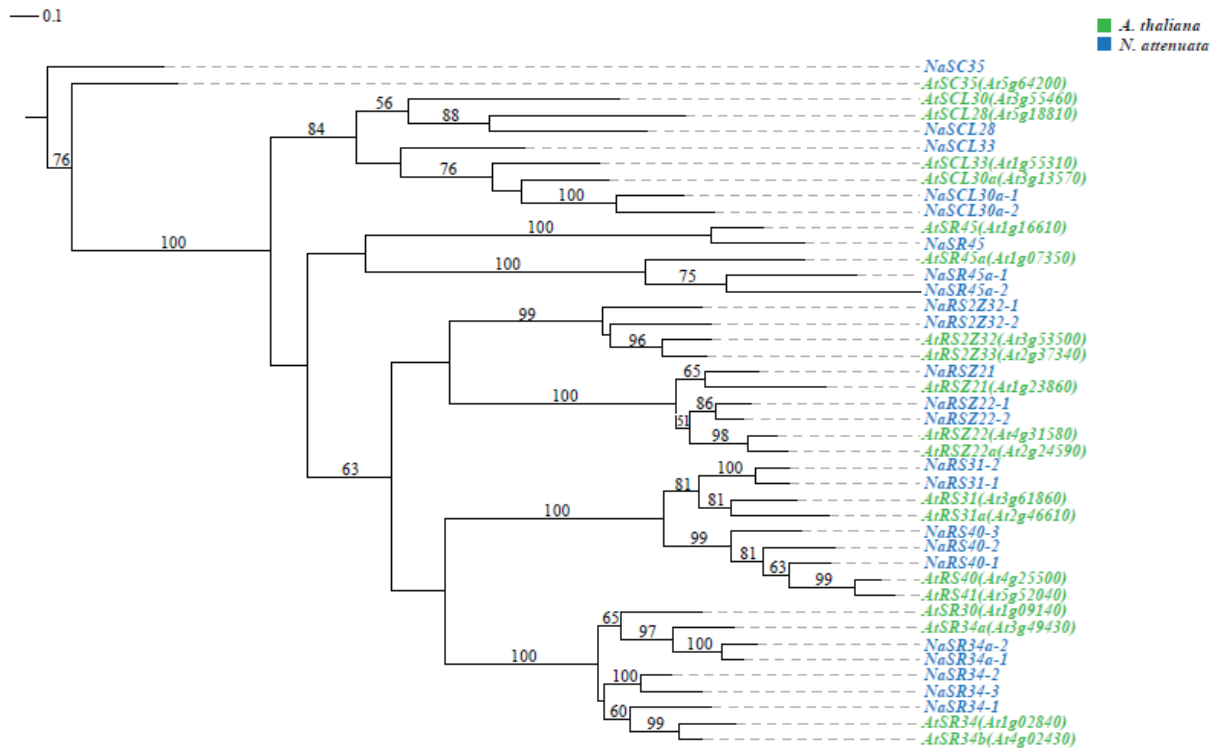


Figure S8. Phylogenetic tree of *A. thaliana* and *N. attenuata* SR and SR-like genes estimated using maximum likelihood method. *A. thaliana* and *N. attenuata* genes are shown in green and blue colours, respectively. Bootstrap values greater than 50 were shown above each branch.

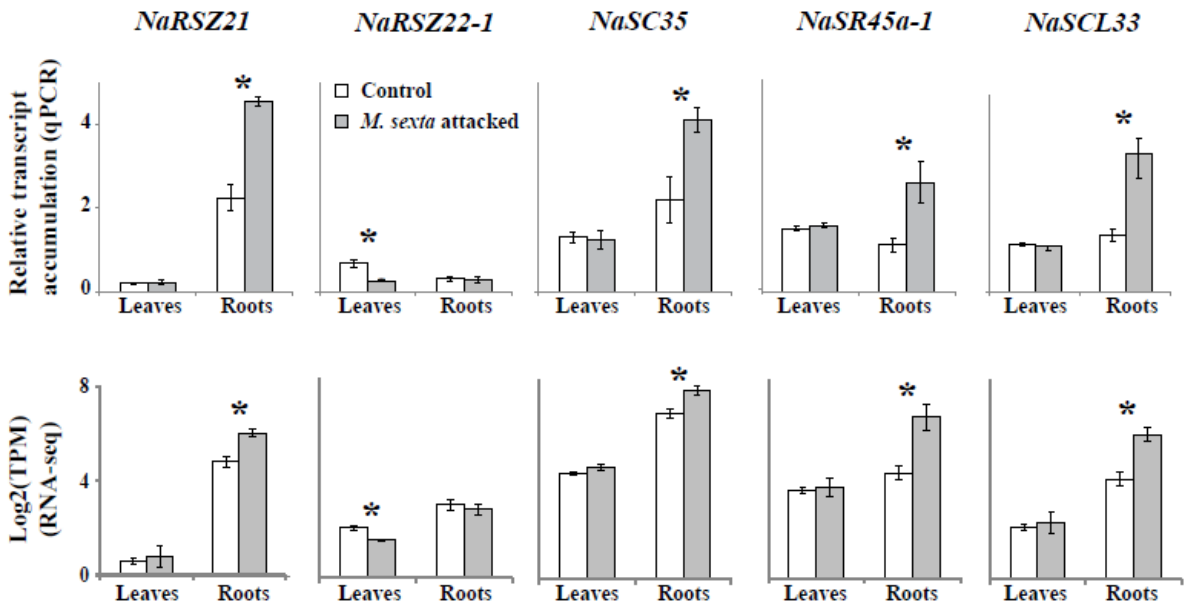


Figure S9. The expression of five selected SR gene expression in leaves and roots measured by RNA-seq and qPCR. The above and below panels show the expression results estimated by qPCR and RNA-seq, respectively. Y-axis in each panel refers to the relative transcript expression level. White bars refer to control and grey bars refer to 5 h of *M. sexta* feeding. TPM refers to transcripts per million. The asterisk in both panels indicates the significance determined by Student's-*t* test ($P < 0.05$). Error bars refer to standard error (SE).

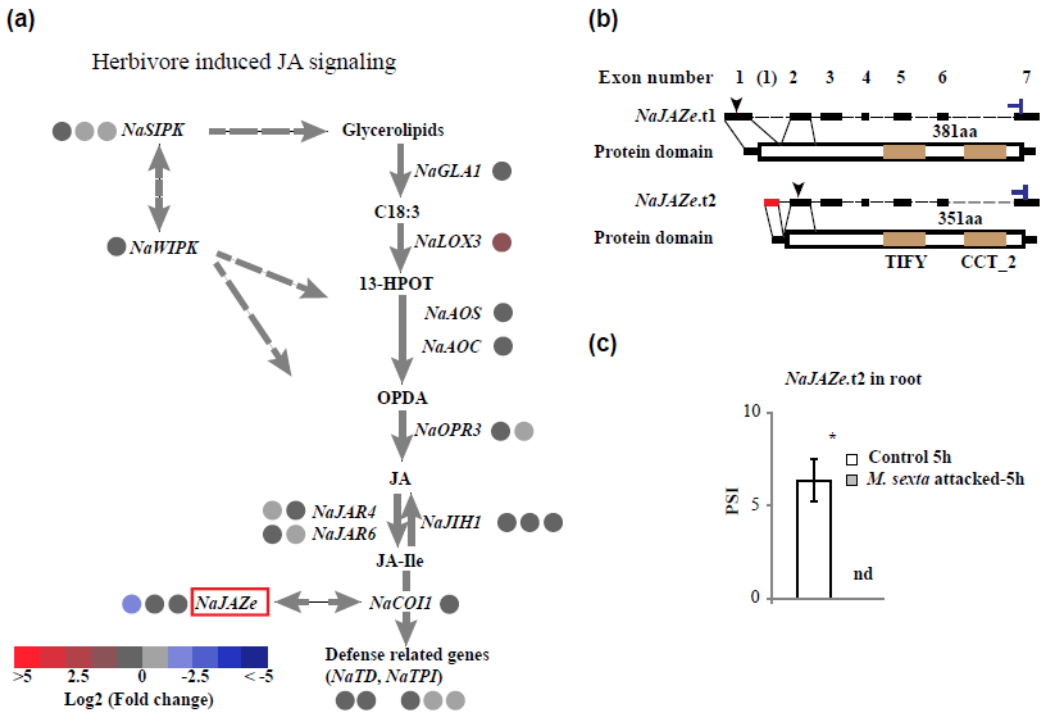


Figure S10. *M. sexta* attack induced transcripts expression changes and AS responses on roots jasmonic acid (JA) signal pathway. (a), *M. sexta* attack (5 h) induced transcripts expression changes on genes involved in JA signalling pathway in root. Each filled circle refers to one transcript with colour indicates the $\log_2$ fold change in comparison to control. Red rectangle box depicts the differentially spliced event induced by *M. sexta* attack. *NaJIH1*, jasmonoyl-l-isoleucine hydrolase 1; *NaSIPK*, salicylate-induced protein kinase; *NaWIPK*, wound-induced protein kinase; *NaGLA1*, glycerolipase 1; *NaLOX3*, lipoxygenase 3; *NaAOS*, allene oxide synthase; *NaAOC*, allene oxide cyclase; *NaOPR3*, OPDA reductase 3; *NaJAR4/6*, jasmonate-resistant 4/6; *NaCOI1*, coronatine-insensitive protein 1; *NaJAZe*, a member of jasmonate ZIM-domain protein family; *NaTD*, threonine deaminase; *NaTPI*, trypsin proteinase inhibitor; **(b)**, the structure of the two *NaJAZe* transcripts resulted from *M. sexta* attack induced AS event in *NaJAZe*. The exon involved in the AS event is shown in red colour. The predicted protein domains are shown in brown colour. The start and stop codons are indicated with black arrow and stop sign, respectively; **(c)**, the predicted percentage splicing index (PSI) of the minor transcript calculated from RNA-seq data; White and grey bars refer to control and attacked samples, respectively. nd refers to not detected. The asterisk indicates the significance that was determined by Students'-*t* test ($P < 0.05$). The error bar refers to standard error (SE).

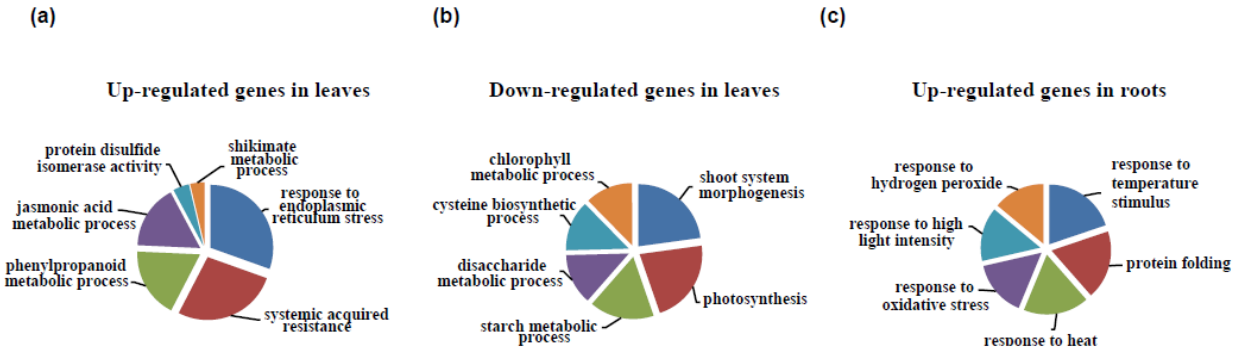


Figure S11. The enriched gene ontology (GO) terms of up and down regulated genes in *N. attenuata* leaves and roots after 5 h of *M. sexta* feeding. (a) and **(b)**, GO terms enriched in up- and down-regulated genes in leaves; **(c)**, GO terms enriched in up-regulated genes in roots.

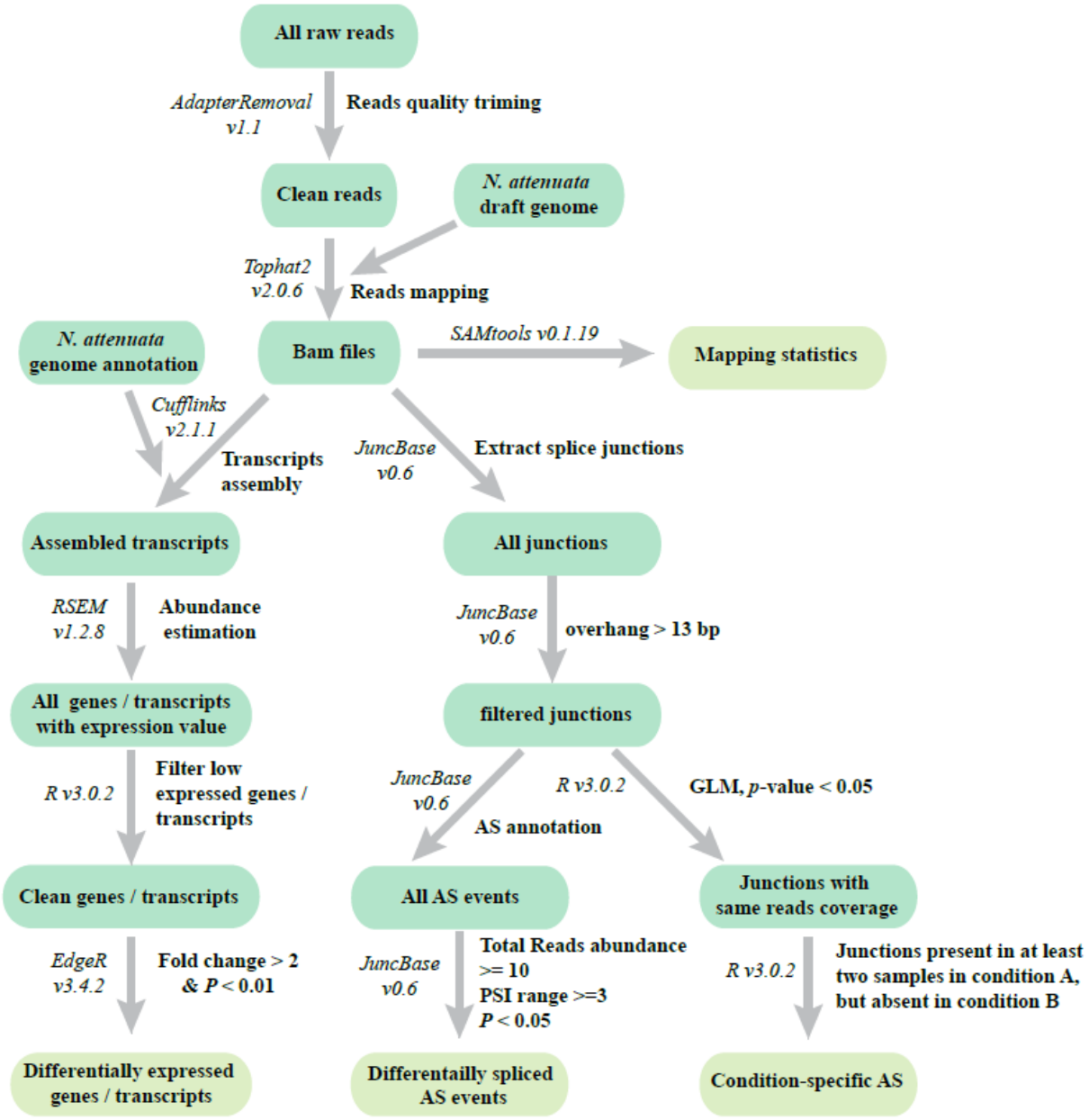


Figure S12. The flowchart of the data analysis process. The software used and bioinformatics processes are shown on the left and right side of each arrow respectively.

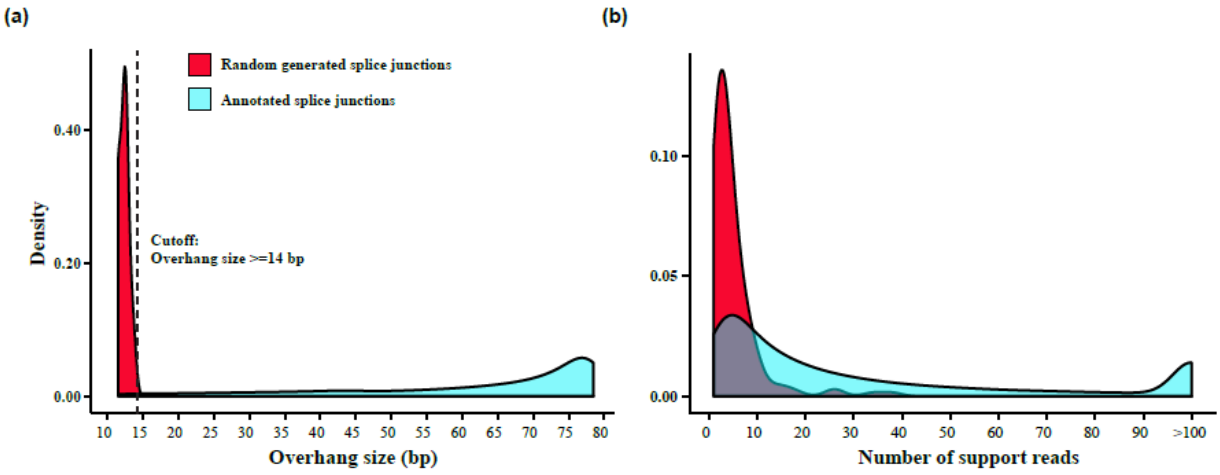


Figure S13. Overhang size and reads coverage distributions in positive and negative SJs datasets. Red colour refers to negative SJs dataset and light blue colour refers to positive SJs dataset. **(a)**, the distribution of the overhang size in negative (random generated junctions) and positive (annotated junctions) splicing junction (SJs) datasets. Y-axis refers to distribution density; X-axis refers to overhang size. While more than 94% positive SJs have overhang size greater than 14 bp, none of the negative SJs had an overhang size greater than 14 bp, which was used as selecting criteria in our downstream analysis (indicated by dashed line); **(b)**, the distribution of read coverage in negative and positive SJs datasets. Y-axis refers to distribution density; X-axis refers to number of supporting reads. Both positive and negative datasets have a large proportion of SJs with low read coverage.

Supplemental tables

Table S1: Mapping results of RNA-sequencing data. The library id was named as following: project ID (ZL), treatment (C or MS refer to control and *M. sexta* feeding, respectively), time points (5 h or 9 h), replicate number (01 to 03) and tissue (L and R refer to leaves and roots, respectively).

Tissue	Time point	Treatment	Library ID	# sequenced reads	# trimmed reads	# mapped reads	% mapped reads	# uniquely mapped reads	% uniquely mapped reads	# reads uniquely mapped to junctions	% reads uniquely mapped to junctions
Leaves	5 hours	Control	ZLC501L	48,354,978	42,816,970	38,003,658	88.76	35,988,665	84.05	9,583,819	22.38
			ZLC502L	48,671,122	35,439,780	31,932,990	90.10	30,214,099	85.25	8,103,948	22.87
			ZLC503L	48,094,454	41,849,942	38,122,667	91.09	36,132,133	86.34	9,740,714	23.28
		M. sexta feeding	ZLMS501L	48,605,098	35,837,738	32,813,963	91.56	31,067,176	86.69	8,354,767	23.31
			ZLMS502L	47,828,114	43,455,662	39,830,285	91.66	37,657,025	86.66	10,089,159	23.22
			ZLMS503L	48,057,710	43,049,656	37,430,302	86.95	35,345,480	82.10	9,161,884	21.28
	9 hours	*M. sexta* feeding	ZLMS901L	48,434,698	45,042,228	39,694,826	88.13	37,103,113	82.37	10,217,300	22.68
			ZLMS902L	48,553,724	40,802,656	36,664,925	89.86	34,104,221	83.58	9,370,437	22.97
			ZLMS903L	49,039,874	27,837,292	24,686,719	88.68	23,250,001	83.52	6,482,680	23.29
Roots	5 hours	Control	ZLC501R	48,169,578	35,607,560	31,956,353	89.75	30,809,154	86.52	8,413,974	23.63
			ZLC502R	48,961,176	35,892,794	32,114,527	89.47	31,093,419	86.63	8,463,855	23.58
			ZLC503R	48,161,738	41,318,960	37,546,618	90.87	36,315,173	87.89	9,929,852	24.03
		M. sexta feeding	ZLMS501R	48,577,352	45,381,648	39,769,697	87.63	38,495,884	84.83	10,243,609	22.57
			ZLMS502R	47,738,032	39,404,020	35,891,213	91.09	34,731,114	88.14	9,723,952	24.68
			ZLMS503R	49,952,308	42,740,750	38,144,505	89.25	36,883,577	86.30	9,783,108	22.89
	9 hours	*M. sexta* feeding	ZLMS901R	50,005,654	43,545,940	39,766,763	91.32	38,532,309	88.49	10,524,775	24.17
			ZLMS902R	48,739,748	38,969,934	34,755,746	89.19	33,586,773	86.19	9,197,504	23.60
			ZLMS903R	48,856,514	43,047,648	38,928,537	90.43	37,584,624	87.31	10,234,457	23.77

Table S2. Total number of annotated AS events and number of genes involved in identified AS events.

Tissue	Control at 5 h		5 h of *M. sexta* feeding		9 h of *M. sexta* feeding	
	# of AS events	# of genes	# of AS events	# of genes	# of AS events	# of genes
Leaves	24063	8500	23182	8247	23968	8474
Roots	29179	9838	32214	10347	31700	10282
Leaves and roots combined	37217	11282	38917	11483	38600	11439

Table S3. Expression of transcripts involved in 5 h *M. sexta* attack induced differentially spliced (DS) genes measured by qPCR. The gene name, Gene ID and transcript ID are shown on left side. t1 and t2 refer to dominant and minor transcripts, respectively. The expression values are shown as average expression value ± standard errors (SE). *P*-values and fold changes were determined by comparing control and *M. sexta* attacked samples (Student's-*t* test.). Transcripts that showed significant ($P < 0.05$) up-regulations are indicated in red colour.

Gene name	Gene ID	Transcript	Leaf			
			Control at five hours	Five hours *M. sexta* feeding	*P*-value	fold change
NaEPSP1	NIATv7_g05371	t1	0.445±0.056	0.992±0.099	0.009	2.231
		t2	0.030±0.004	0.034±0.008	0.344	1.129
NaPPDK	NIATv7_g29964	t1	7.137±0.604	10.738±2.051	0.087	1.505
		t2	0.043±0.023	0.328±0.091	0.039	7.644
NaXCT	NIATv7_g031395	t1	2.324±0.135	2.107±0.044	0.100	0.906
		t2	0.091±0.008	0.168±0.029	0.031	1.844
			Root			
NaERD6	NIATv7_g27851	t1	0.903±0.049	0.971±0.033	0.403	1.076
		t2	0.070±0.016	0.136±0.017	0.023	1.937
NaU2AF65B	NIATv7_g01580	t1	0.430±0.099	0.241±0.002	0.117	0.560
		t2	0.042±0.012	0.029±0.0001	0.221	0.683

Table S4. **The expression of all 23 SR and SR-like genes in leaves and roots.** The sub-family names, gene ID, gene names are shown on left side. The expression values were shown as $\log_2$ (TPM). *P*-values and fold changes were determined by comparing control and *M. sexta* attacked samples. Significantly ($P < 0.05$) up and down-regulations are indicated in red and green, respectively.

Gene name	Gene ID	Leaf											Root									
		Control at five hours			Five hours *M. sexta* feeding			Nine hours *M. sexta* feeding			*P*-value	fold change	Control at five hours			Five hours *M. sexta* feeding			Nine hours *M. sexta* feeding			*P*-value	fold change
Sample ID		ZLC50 1L	ZLC50 2L	ZLC50 3L	ZLMS50 1L	ZLMS50 2L	ZLMS50 3L	ZLMS90 1L	ZLMS90 2L	ZLMS90 3L			ZLC50 1R	ZLC50 2R	ZLC50 3R	ZLMS50 1R	ZLMS50 2R	ZLMS50 3R	ZLMS90 1R	ZLMS90 2R	ZLMS90 3R		
NaSR34-like1	NIATv7_g27565	2.28	1.65	2.18	1.58	1.97	2.61	2.41	1.69	1.71	0.48	1.01	1.63	2.13	2.10	2.07	2.30	2.37	2.63	1.99	2.06	0.09	1.23
NaSR34-like2	**NIATv7_g05821**	4.34	4.14	4.42	4.20	4.02	3.96	4.31	3.98	4.21	0.05	0.84	6.07	6.18	6.23	5.65	6.09	5.55	6.22	6.16	6.04	**0.04**	**0.76**
NaSR34-like3	**NIATv7_g38046**	5.50	5.32	5.50	5.33	5.30	4.92	5.77	5.75	5.93	0.08	0.84	6.02	6.05	6.07	5.35	5.59	5.29	5.93	5.91	5.94	**0.00**	**0.64**
NaSR34a-like1	NIATv7_g04107	3.00	2.50	2.89	2.65	2.88	2.64	2.99	2.99	3.33	0.34	0.95	4.23	4.02	4.19	4.14	3.84	3.83	4.04	3.87	4.26	0.08	0.87
NaSR34a-like2	**NIATv7_g15184**	2.56	2.34	2.14	2.42	2.67	3.15	3.15	3.14	3.03	0.09	1.32	4.68	4.29	4.62	4.74	5.05	4.97	5.04	4.92	4.77	**0.03**	**1.31**
NaRSZ21	**NIATv7_g00053**	0.85	0.43	0.53	0.15	0.53	1.70	0.57	1.21	0.88	0.36	1.14	5.22	4.43	4.72	6.01	5.75	6.28	4.65	4.49	4.14	**0.01**	**2.33**
NaRSZ22-like1	**NIATv7_g12462**	2.01	1.98	2.13	1.52	1.46	1.52	2.30	1.94	1.85	**0.00**	**0.69**	2.66	2.88	3.26	2.63	3.09	2.48	3.28	3.00	3.17	0.24	0.87
NaRSZ22-like2	NIATv7_g30927	4.49	4.25	4.49	4.26	4.17	4.52	4.40	4.42	4.52	0.25	0.94	5.36	5.44	5.43	5.18	5.54	5.34	5.32	5.25	5.24	0.31	0.96
NaSC35	**NIATv7_g00887**	4.11	4.32	4.30	4.38	4.31	4.69	4.63	4.72	4.55	0.09	1.16	6.93	6.31	6.72	7.78	7.22	7.70	6.89	6.83	6.71	**0.01**	**1.88**
NaSCL28	**NIATv7_g29093**	2.27	2.13	2.17	1.84	1.32	1.74	2.16	2.01	1.66	**0.01**	**0.68**	2.47	2.89	2.98	2.72	2.56	2.66	3.34	2.74	2.90	0.23	0.91
NaSCL30a-like1	**NIATv7_g31388**	4.57	4.50	4.61	4.34	4.45	5.07	5.02	5.12	5.02	0.41	1.04	6.04	6.00	6.02	6.24	6.24	6.41	6.31	6.28	6.17	**0.00**	**1.21**
NaSCL30a-like2	**NIATv7_g28876**	4.20	3.81	4.14	4.13	4.20	3.88	4.24	4.08	4.26	0.45	1.02	5.42	5.67	5.55	4.97	4.91	4.94	4.88	5.19	5.28	**0.00**	**0.66**
NaSCL33	**NIATv7_g29904**	1.82	2.21	2.14	1.54	2.20	3.07	2.63	2.99	2.96	0.33	1.16	4.57	3.72	3.73	5.94	5.23	6.20	4.60	4.34	3.74	**0.01**	**3.45**
NaRS2Z32-like1	NIATv7_g36741	5.27	4.90	5.40	5.19	5.07	5.15	5.23	5.06	5.10	0.38	0.96	6.53	6.50	6.42	6.25	6.44	6.37	6.44	6.43	6.30	0.05	0.87
NaRS2Z32-like2	NIATv7_g07938	4.25	4.09	4.02	4.03	3.86	3.91	4.24	4.52	4.11	0.05	0.88	5.79	6.15	5.93	5.77	5.88	5.54	6.22	6.23	6.07	0.09	0.85
NaRS31-like1	**NIATv7_g33359**	1.45	1.04	1.12	1.10	1.45	1.30	1.46	1.40	1.13	0.32	1.06	3.25	2.81	2.94	3.56	3.63	3.71	4.13	3.67	3.80	**0.01**	**1.55**
NaRS31-like2	NIATv7_g36650	3.19	3.01	3.22	3.00	3.02	2.45	3.19	2.89	2.93	0.09	0.80	5.11	4.73	4.89	4.91	5.00	4.71	5.15	5.03	4.86	0.39	0.97
NaRS40-like1	NIATv7_g21300	5.53	5.46	5.48	5.64	5.52	5.68	5.51	5.49	5.79	0.05	1.09	5.97	6.29	6.34	5.90	6.20	5.82	6.31	6.20	6.20	0.12	0.85
NaRS40-like2	NIATv7_g09436	5.31	5.07	5.21	5.37	5.25	5.52	5.23	5.43	5.28	0.08	1.14	6.28	6.33	6.39	6.09	6.32	5.94	6.67	6.41	6.45	0.07	0.86
NaRS40-like3	NIATv7_g24643	2.14	2.12	2.40	2.47	2.12	2.56	2.81	3.05	2.28	0.18	1.12	4.02	3.74	4.02	4.01	4.15	3.83	4.67	4.46	4.41	0.30	1.05
NaSR45	NIATv7_g08524	5.24	5.08	5.46	5.51	5.14	4.99	5.44	5.29	5.28	0.41	0.97	7.68	7.29	7.46	7.16	7.35	7.19	7.56	7.52	7.31	0.06	0.85
NaSR45a-like1	**NIATv7_g21819**	3.51	3.34	3.72	3.11	3.48	4.33	3.27	3.49	3.00	0.39	1.08	4.80	4.05	3.89	6.60	5.51	7.32	3.93	4.24	3.85	**0.01**	**4.69**

Table S5. Table listing the percentage splicing indexes (PSI) of all AS events identified in 23 SR and SR-like genes. Gene family name, gene ID, genomic locations and AS types are shown on left side, PSI and *P*-values are shown on right side. *P*-values were determined by comparing control and 5 h *M. sexta* attacked samples (Student's-*t* test). NA refers to the SJs was found not expressed in the sample.

Subfamily	Gene name	GeneID	Related genomic coordinates 1	Related genomic coordinates 2	Strand	Event ID	AS type	Leaf							Root						
								Control five hours			Five hours *M. sexta* feeding			*p*-value	Control five hours			Five hours *M. sexta* feeding			*p*-value
Sample ID								ZLC501L	ZLC502L	ZLC503L	ZLMS501L	ZLMS502L	ZLMS503L		ZLC501R	ZLC502R	ZLC503R	ZLMS501R	ZLMS502R	ZLMS503R	
SR	*NaSR34-like1*	NIATv7_g27565	scaffold004527:84362-87217	84430-84895	-	AS1	ES	36.4	44.4	50.0	28.6	28.6	42.9	0.2	0.0	0.0	0.0	0.0	0.0	50.0	0.4
			scaffold004527:87491-87571	-		AS2	IR	10.0	0.0	0.0	25.0	0.0	28.6	0.5	0.0	20.0	0.0	0.0	0.0	NA	0.4
	NaSR34-like2	NIATv7_g05821	scaffold000553:188757-189143	188757-189130	+	AS1	AltA	0.0	10.0	9.1	NA	0.0	0.0	0.2	3.5	1.5	2.4	6.9	3.2	7.1	0.1
			scaffold000553:194224-194396	-		AS2	IR	2.0	4.3	7.6	2.2	7.0	6.5	0.8	0.8	3.9	3.4	4.6	2.6	4.4	0.4
			scaffold000553:193739-193823	-		AS3	IR	23.5	31.3	31.3	18.2	30.8	30.0	0.7	NA	18.8	20.0	0.0	17.7	NA	0.4
			scaffold000553:192486-192571	192486-192568		AS4	AltA	0.0	0.0	0.0	3.3	0.0	0.0	0.4	1.6	0.0	0.0	0.0	1.2	0.0	0.8
	NaSR34-like3	NIATv7_g38046	scaffold008651:25566-27278	-	+	AS1	IR	15.7	14.1	16.5	11.4	13.9	25.0	0.8	5.1	5.8	4.1	7.1	8.8	4.0	0.4
			scaffold008651:25566-27278	26743-26827		AS2	ES	4.2	4.3	2.6	1.4	3.1	2.7	0.2	0.7	1.5	0.7	0.0	1.1	1.4	0.8
			scaffold008651:25179-25269	-		AS3	IR	3.3	0.0	1.6	1.7	0.0	7.9	0.6	1.6	4.2	7.0	7.0	4.1	4.8	0.6
			scaffold008651:25377-25466	25377-25463		AS4	AltA	0.0	1.6	0.0	0.0	0.0	0.0	0.4	2.0	1.0	2.3	1.6	0.0	2.0	0.5
	NaSR34a-like1	NIATv7_g04107	scaffold000393:246345-249069	248652-248976	+	AS1	ES	NA	27.3	13.3	8.3	7.1	37.5	0.8	7.3	17.8	18.5	15.8	15.2	12.1	1.0
	NaSR34a-like2	NIATv7_g15184	scaffold001879:145682-147270	-	+	AS1	IR	NA	NA	NA	NA	NA	NA	NA	0.0	5.6	7.1	4.6	3.0	2.4	0.7
RSZ	*NaRSZ21*	NIATv7_g00053	scaffold000013:494896-497119	494896-495105	+	AS1	AltA	NA	NA	NA	NA	NA	NA	NA	0.0	2.9	0.0	0.0	1.1	0.7	0.7
	NaRSZ22-like1	NIATv7_g12452	-	-	-	nd	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NaRSZ22-like2	NIATv7_g30927	scaffold005501:45519-45723	-	+	AS1	IR	93.8	100.0	77.3	94.7	100.0	95.2	0.5	100.0	100.0	91.7	95.5	100.0	100.0	0.7
SC	*NaSC35*	NIATv7_g00887	scaffold000057:243196-243298	243217-243298	-	AS1	AltA	NA	NA	NA	NA	NA	NA	NA	99.3	100.0	99.1	99.5	100.0	95.2	0.5
SCL	*NaSCL28*	NIATv7_g29093	scaffold004939:56099-57781	56950-57781	-	AS1	AltA	66.7	50.0	70.0	50.0	62.5	100.0	0.6	50.0	72.7	54.6	NA	NA	76.9	NA
	NaSCL30a-like1	NIATv7_g31385	scaffold005670:26998-27929	27412-27555	+	AS1	ES	12.5	7.9	13.9	16.1	12.5	17.5	0.2	7.4	12.0	12.2	11.0	14.0	13.4	0.3
			scaffold005670:26998-27929	-		AS2	IR	9.7	5.4	6.1	13.3	9.7	8.3	0.2	7.4	6.9	7.7	7.3	7.5	5.8	0.5
	NaSCL30a-like2	**NIATv7_g28876**	scaffold004894:34637-35592	34981-35198	+	**AS1**	**ES**	7.1	NA	6.3	7.1	20.0	NA	0.5	4.2	2.4	5.6	9.7	14.3	13.0	**0.0**
			scaffold004894:34637-35592	-		AS2	IR	NA	NA	NA	NA	NA	NA	NA	0.0	7.0	8.1	3.5	10.0	16.7	0.3

Table S6: The accession number of genes that were selected for AS and gene expression analysis.

Gene name	GeneID	Accession ID
NaBAM2	NIATv7_g03670	GCZL01000001
NaESP3	NIATv7_g17815	GCZL01000002
NaEPSP1	NIATv7_g05371	GCZL01000003
NaPPDK	NIATv7_g29964	GCZL01000004
NaXCT	NIATv7_g31395	GCZL01000005
NaERD6	NIATv7_g27851	GCZL01000006
NaU2AF65B	NIATv7_g01580	GCZL01000007
NaSR34-like1	NIATv7_g27565	GCZL01000008
NaSR34-like2	NIATv7_g05821	GCZL01000009
NaSR34-like3	NIATv7_g38046	GCZL01000010
NaSR34a-like1	NIATv7_g04107	GCZL01000011
NaSR34a-like2	NIATv7_g15184	GCZL01000012
NaRSZ21	NIATv7_g00053	GCZL01000013
NaRSZ22-like1	NIATv7_g12452	GCZL01000014
NaRSZ22-like2	NIATv7_g30927	GCZL01000015
NaSC35	NIATv7_g00887	GCZL01000016
NaSCL28	NIATv7_g29093	GCZL01000017
NaSCL30a-like1	NIATv7_g31385	GCZL01000018
NaSCL30a-like2	NIATv7_g28876	GCZL01000019
NaSCL33	NIATv7_g29904	GCZL01000020
NaRS2Z32-like1	NIATv7_g36741	GCZL01000021
NaRS2Z32-like2	NIATv7_g07938	GCZL01000022
NaRS31-like1	NIATv7_g33359	GCZL01000023
NaRS31-like2	NIATv7_g36650	GCZL01000024
NaRS40-like1	NIATv7_g21300	GCZL01000025
NaRS40-like2	NIATv7_g09436	GCZL01000026
NaRS40-like3	NIATv7_g24643	GCZL01000027
NaSR45	NIATv7_g08524	GCZL01000028
NaSR45a-like1	NIATv7_g21819	GCZL01000029
NaSR45a-like2	NIATv7_g11571	GCZL01000030

Table S7: The primers used for validating differentially expressed SR genes and differentially spliced transcripts.

Gene ID	Primer ID	Forward primer	Reverse primer	Transcript ID	Note	Primer Locations
NIATv7_g05371	9756	GGACAGAGAACAGCGTCACA	GCGCTCAGTTTCCTTGACTC	*NaEPSP1.t1*	primers used for measuring the expression of *NaEPSP1.t1* transcript that contains the skipped Exon	F: the 6th exon/ R: the 8th exon
NIATv7_g05371	9757	GCAGTTTACAGTCGCTAGCT	TGGCCATTCTGTGATCGTCG	*NaEPSP1.t2*	primers used for measuring the expression of *NaEPSP1.t2* transcript that skiped the exon.	F: span the junction of 5th and 7th exons/ R: the 8th exon
NIATv7_g29964	52699	CAACAAAGGGGAGTTTTTTTG	CCCATTTGCTTCTTGACTGC	*NaPPDK.t1*	Primers used for measuring the expression of *NaPPDK.t1* transcript that has regular 2nd exon	F: the longer part of 2nd exon / R: the 3rd exon
NIATv7_g29964	52701	CTCAATCACAACAAAGGATTTGA	TGAGGTGGAAGAGAGCCAAT	*NaPPDK.t2*	Primers used for measuring the expression of *NaPPDK.t2* transcript that has Alt3' (20bp) in the 2nd exon	F: the common part of 2nd exon shared by both transcripts / R: the 3rd exon
NIATv7_g31395	50552	ACCCTCGGCTATCTTTTTGC	CCCTCTCACTGTCCGGTAAG	*NaYCT.t1*	Primers used for measuring the expression of *NaYCT.t1* transcript that has short exon	F: the 4th exon/ R: span the junction of 5th and shorter 6th exons
NIATv7_g31395	50550	ACCCTCGGCTATCTTTTTGC	TCCCTCTCACTGAAAAAGTG	*NaYCT.t2*	Primers used for measuring the expression of *NaYCT.t2* transcript that has long exon	F: the 4th exon/ R: the longer part of 6th exons
NIATv7_g27851	46239	TCATGTGTGGGATACTCAGCA	ATTCTCATGGCCCCTTTTCT	*NaERD6.t1*	Primers used for measuring the expression of *NaERD6.t1* transcript that has short 2nd exon	F: span the junction of 1st and 2nd shorter exons/ R: the 3rd exon
NIATv7_g27851	49240	TTGGATCATGTGACCATGAAA	CAATTCTGCCGCTTGTGATA	*NaERD6.t2*	Primers used for measuring the expression of *NaERD6.t2* transcript that has long 2nd exon	F: the longer part of 2nd exon / R: the 3rd exon
NIATv7_g01580	2907	CTGCTGTTACAGGTCAACTTCC	ACCCACATAAACTCGCCTTG	*NaU2AF65B.t1*	Primers used for measuring the expression of *NaU2AF65B.t1* transcript that skips the exon	F: span the junction of 5th and 7th exons/ R: the 8th exon
NIATv7_g01580	2906	ACTCGAGGGGAAAATCTGGT	TCAGCACAGTGGGGCTACTA	*NaU2AF65B.t2*	Primers used for measuring the expression of *NaU2AF65B.t2* transcript that includes the exon	F: the 4th exon/ R: the 6th exon
NIATv7_g29964	61625	ATGCCGCTATTCCACAACAT	GGAATTTAGCAACTGTGACT	*NaJIH1.t1*	Primers used for measuring the expression of the *NaJIH1.t1* transcript that includes the exon	F: the 3rd exon/ R: the 4th exon
NIATv7_g29964	61624	TTTTGGCTGGAAGTGGTTTT	CCCAACAATAACCTGCGAAT	*NaJIH1.t2*	Primers used for measuring the expression of the *NaJIH1.t2* transcript that skips the exon	F: the 3rd exon/ R: span the junction of 3th and 5th exons
NIATv7_g00053	RSZ21	GGCCGCTATGTGTCACCTTA	TGAAGAGCCTCAGTTGTATTTCC	*NaRSZ21*	Primers used for measuring the expression of *NaRSZ21*	F: the 5th exon/ R: the 7th exon
NIATv7_g12452	RSZ22	ATTTTCTGGCTGCTTCTTTT	CCAGCTCTTTCTCCAAAGACA	*NaRSZ22-1*	Primers used for measuring the expression of *NaRSZ22-1*	F and R: the 7th exon
NIATv7_g00887	SC35-2	GGCTAATGAAGTCCGTCCAC	TGACTCATTTAAGTTCCCAACCA	*NaSC35*	Primers used for measuring the expression of *NaSC35*	F and R: the 7th exon
NIATv7_g21819	SR45a	TCCCGGGATGATCGTTATTA	CCGCCTAGGTGATACACTGC	*NaSR45a-1*	Primers used for measuring the expression of *NaSR45a-1*	F and R: the 7th exon
NIATv7_g29904	SCL33	ATTCTCGATCTCCACGGTGT	TTCTGCTCCTAGAACCATAAGGA	*NaSCL33*	Primers used for measuring the expression of *NaSCL33*	F and R: the 5th exon

Manuscript II

Wild tobacco genomes reveal the evolution of prolific nicotine production

Shuqing Xu, Thomas Brockmöller, Aura Navarro-Quezada, Heiner Kuhl, Klaus Gase, Zhihao Ling, Wenwu Zhou, Christoph Kreitzer, Mario Stanke, Haibao Tang, Eric Lyons, Priyanka Pandey, Shree P. Pandey, Bernd Timmermann, Emmanuel Gaquerel, and Ian T. Baldwin

Submitted to Nature Plants (10.2016)

Wild tobacco genomes reveal the evolution of prolific nicotine production

Shuqing Xu[1,#*], Thomas Brockmöller[1#], Aura Navarro-Quezada[2], Heiner Kuhl[3], Klaus Gase[1], Zhihao Ling[1], Wenwu Zhou[1], Christoph Kreitzer[1,4], Mario Stanke[5], Haibao Tang[6], Eric Lyons[7], Priyanka Pandey[8], Shree P. Pandey[9], Bernd Timmermann[3], Emmanuel Gaquerel[2*], and Ian T. Baldwin[1*]

Affiliations:

[1]Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Jena, Germany.

[2]Centre for Organismal Studies, University of Heidelberg, Heidelberg, Germany.

[3]Max Planck Institute for Molecular Genetics, Berlin, Germany.

[4]Veterinary University of Vienna, Vienna, Austria.

[5]Institute for Mathematics and Computer Science, Universität Greifswald, Greifswald, Germany.

[6]Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China.

[7]School of Plant Sciences, BIO5 Institute, CyVerse, University of Arizona, Tucson, USA.

[8]National Institute of Biomedical Genomics, Kalyani, West Bengal, India.

[9]Department of Biological Sciences, Indian Institute of Science Education and Research - Kolkata, 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

Plants adapt to their environments by evolving novel traits, a process intimately linked and reflected in their remarkable capacity to produce structurally diverse specialized metabolites. A majority of these metabolites function as multifunctional chemical shields against a variety of biotic and abiotic stresses [1,2]. 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 the *Nicotiana* genomes, promoted expression divergences among duplicated genes and contributed to the evolution of herbivory-induced signaling and defenses, including nicotine biosynthesis, in these species. The biosynthetic machinery allowing for nicotine synthesis in the roots [3] 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 capable of producing nicotine, albeit in trace quantities, the prolific production and ecological utilization of this toxin in *Nicotiana* required lineage-specific duplications within the NAD pathway and a retro-transposon mediated duplication of *berberine bridge enzyme-like* involved in the coupling of the two pathways. Furthermore, TE insertions that incorporated transcription factor binding motifs 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.

Main text

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 that feed on tobacco plants. 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 [4-6]. 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 [6]. 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 [7]. 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 [8] – *i.e.* up to 30,000 times higher than in other *Solanaceae* genera. This prodigious biosynthetic ability is based on an efficient biosynthetic machinery composed of multiple genes co-expressed in roots and which commandeers a substantial fraction of the plant's nitrogen assimilation [3,9,10]. In some species, such as *N. obtusifolia*, a sympatric perennial relative of *N. attenuata,* nicotine is rapidly demethylated at the root-shoot junction to form nornicotine, perhaps as part of a species-specific self-toxicity mitigating mechanism [11]. In contrast to the deep knowledge that exists on nicotine's biosynthesis and ecological functions, the evolutionary origin and genomic features that facilitated the assembly of a pathway so critical for the survival of *Nicotiana* species have 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 [12,13]. 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 [12-17]. Hence, the combination of gene duplications and TE activity is thought to facilitate the evolution of novel adaptive traits [13]. However, the details of this process, in particular its role in the evolution of plant metabolic complexity through the assembly of novel multi-gene pathways, remains unclear.

We sequenced and assembled the genome of *N. attenuata* (accession Utah, collected in the Southwestern Utah in 1996), using 30x Illumina short reads, 4.5x 454 reads, and 10x PacBio single-molecule long reads. In total, 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 (Supplementary Figure 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 (Supplementary Figure 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. For *N. attenuata*, more than 99.6% of whole-genome shotgun reads and 96.5% of RNA-seq reads can be mapped to the genome assembly. Analyses of 248 conserved core eukaryotic genes using the CEGMA[18] pipeline indicated that both assemblies were only slightly less complete in full-length gene contents than that of the tomato genome (86.7%) and similar to the potato genome (83.9%) (Supplementary Table S2).

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 validate the gene model predictions, 428 selected genes which has been the subject of functional analysis in *N. attenuata* were manually inspected and manually curated, and pseudogenes were excluded. Because gene models within a genus are usually conserved, we annotated *N. obtusifolia* and two other publicly available diploid *Nicotiana* genomes [19] using a homology-based annotation pipeline based on all predicted *N. attenuata* protein coding gene

models. In total, 27,911, 27,724 and 26,455 full-length protein coding genes were annotated in *N. obtusifolia, N. sylvestris* and *N. tomentosiformis,* respectively.

To investigate the evolutionary history of the different *Nicotiana* genomes, we inferred 23,340 homologous groups using protein sequences from 11 published genomes (Supplementary Table S5). 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*, but not with *Mimulus* (Figure 1, Supplementary Figure S13-16). 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. Among other previously characterized herbivory-related genes, *THREONINE DEAMINASE* (*TD*) is an example of this trend (Supplementary Figure S14). 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 [20], evolved as an anti-digestive defense in tomato that degrades threonine in the guts of attacking larvae [21]. Interesting, in *N. attenuata TD2.2* likely retained its biosynthetic function to supply the Ile essential for jasmonate-mediated defense signaling [22]. 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 (Supplementary Figure S14).

Polyploidization is often associated with a burst of TE activity as a hypothesized consequence of "genomic shock"[23,24]. Indeed, 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 [25], albeit earlier than in *Nicotiana*, indicating that after WGT, the different *Solanaceae* lineages independently experienced 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 major 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 [26,27]. 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. Analyzing insertion positions of this subgroup revealed that DTT-NIC1 copies, similar to other DNA transposons, are significantly enriched within a 1 kb region upstream of genes (Figure 2C). We next explored the impact that DTT-NIC1 insertions had on insect herbivory-induced genome-wide expression responses in *N. attenuata* within a 21 h time course experiment. DTT-NIC1 insertions are significantly enriched within a 1 kb upstream region of genes that are induced by insect herbivory as compared to the genome-wide background ($p < 0.001$), suggesting that insertions of this TE family had shaped herbivory-induced transcriptomic responses elicited in this *Nicotiana* species (Supplementary Figure S4).

Innovations in metabolic and signaling network architecture are thought to result from the rapid rewiring of tissue-level gene expression patterns following duplications events [28,29]. 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 in *Nicotiana* species was a critical determinant of genome-wide re-wirings of gene regulation occurring post-duplication in *Nicotiana* species.

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 berberine bridge enzyme-like (BBL) enzymes [30,31] (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 that is required for the biosynthesis of the signature alkaloids of *Solanaceae* and *Convolvulaceae*, such as tropane, nortropane, and nicotine, are shared between *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 (Supplementary figure S21, S22). 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[32]. The *N-methylputrescine oxidase* (*MPO*) from the polyamine metabolism pathway was derived from diamine oxidase through genome-wide duplications. Bothe copies are retained in *Nicotiana*, *Solanum* and *Petunia* (Supplementary figure S23), presumably to

sustain the flux of *N*-methyl-Δ^1-pyrrolinium required for alkaloid biosynthesis. Duplication patterns of *ODC*, *PMT* and *MPO* duplication therefore support the ancient origin of the ornithine-derived *N*-methyl-Δ^1-pyrrolinium, which is utilized as a common building block for the biosynthesis of most alkaloid groups in the Solanaceae and Convolvulaceae.

In contrast to the relatively ancient origin of pyrrolidine 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 (Supplementary figure S24, S26). Interestingly, the key innovation in the synthesis of pyridine alkaloids in *Nicotiana* species, BBLs, thought to be likely involved in the late oxidation step in nicotine biosynthesis that couples the pyridine and pyrrolidine rings, originated from a *Nicotiana*-specific *Gypsy* retro-transposon-mediated duplication as indicated by their intronless structures (Supplementary figure S27) and genomic locations close to *Gypsy* elements. However, due to lack of genomic information from the closely related genus of *Nicotiana*, it is remain unclear whether duplications of NAD pathway genes and BBL occurred at the ancestor of all genera, in which some species also produces certain quantify of nicotine, such as *Crenidium*, *Cyphanthera* and *Duboisia*.

Tissue-level RNAseq 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 [33]. This tissue-specific expression and herbivory-responsive regulation of nicotine biosynthetic genes relies in part on the presence of two transcription factor binding sites, the GCC and G-box elements in the promotor regions of the genes. Experimental evidence has shown that these two elements are recognized by transcription factors of the ethylene response factor (ERF) subfamily IX and MYC2, respectively, central players in jasmonate signaling[9]. Nicotine biosynthetic genes also 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 was central to the evolution of the coordinated transcriptional regulation required for high-flux nicotine biosynthesis.

To further investigate the origin of these GCC and G-box motifs, we examined the sequence homology of the 150 bp flanking region harboring these GCC and G-box motifs in the *N. attenuata* genome. This analysis demonstrated that paralogs of at least 38.1% and 30.0% of GCC and G-box flanking sequences, respectively (Figure 3C), showed homology to annotated TEs in *N. attenuata* from different families and classes, suggesting that these motif sequences are likely derived from TE insertions. While G-box motifs are derived from different transposons, the majority of GCC motifs are likely derived from *Gypsy* transposons. For example, after WGT, while two duplicated *PMT* genes were both retained in *Solanum*, only one *PMT* was retained in *Nicotiana* and subsequently tandemly duplicated, which resulted in two root-specifically expressed genes, *PMT1.1* and *PMT1.2*. The 2 kb upstream sequences of these two *N. attenuata PMT* genes contain three and four GCC motifs, of which one and three, respectively, were derived from *Gypsy* TE (Figure 3C).

In addition to directly contributing to promoter binding motifs, the abundance of TE fragments found within 2 kb regions of nicotine biosynthesis genes suggests that other mechanisms, such as the introduction of DNA methylation sites, or generation of new target sites of small RNAs may have also contributed to the tissue-specific expression pattern of nicotine biosynthesis genes. For example, the two root-specific expressed *BBLs* contain LTR and DTT-NIC1 within their 2 kb upstream region (Supplementary figure S27), which may have suppressed their expression in other tissues than roots. Future studies that specifically manipulate the TE regions of these *Nicotiana* promoter elements will provide further mechanistic insights into the roles of these TE fragments in the evolution of nicotine biosynthesis.

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, are 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 quantity of nicotine to ward off attackers, it is likely that this 'junk' has inspired innovation essential for their survival [34].

Materials and Methods:

Plant material and DNA preparation

Plants were grown as previously described[35]. The genomic DNA sequenced by 454 and Illumina HiSeq2000 technologies was isolated from late rosette-stage plants using the CTAB-method [36]. 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 [35]. *N. obtusifolia* 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 [37]. High molecular weight genomic DNA used to generate the optical map of *N. attenuata* was isolated 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 using a nuclei based protocol[38].

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* [39], line A-07-457-1, which was transformed with pRESC5AOC [GenBank KX011463] and is impaired in JA biosynthesis ; father: ir*GGPPS* [40], line A-07-230-5, which was transformed with pRESC5GGPPS [GenBank KX011462] and is impaired in the synthesis of 17-hydroxygeranyllinalool diterpene glycosides), both generated from the 22nd inbred generation of the same origin as the *N.*

attenuata plants described above[35]. 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 Supplementary Figure 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[41]. We used the SSPACE v2 scaffolder [42] to further improve scaffolding using the mate-pair data and filled gaps using GapCloser v1.12 [43]. After manual inspections, we found that certain neighboring contigs in the scaffolds still have 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 overlapped contigs by custom scripts.

The assembly scaffolds from short reads were used for gap filling and further scaffolding using PBJelly (v15.8.24) [44] 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% of errors, we performed an additional correction step using short reads with PILON [45]. 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 [46] 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 [47], 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 similar manner to the *N. attenuata* assembly.

Annotation of transposable elements

De novo annotation of repeated elements was performed with RepeatModeler version open-4-0-5 with the parameters (-engine ncbi). We identified 667 consensus repeat sequences (1.3 Mb total size) in

the *N. attenuata* genome. To classify these consensus repeat sequences, additional annotation using TEclass [48] was used 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 "-e ncbi -norna". We further re-annotated transposable elements using the *N. attenuata* repeat library for four *Nicotiana* additional genomes: *N. sylvestris* and *N. tomentosiformis* [19]. To make the results comparable, we used the same approach to *de novo* identify the TE library of *S. lycopersicum* [49] and *S. tuberosum* [50] genomes.

MITEs in *Nicotiana* were annotated in two steps. First, MITE-Hunter [51] was used to find MITE families in the *N. attenuata* genome using default parameters, except "-P 0.2". Following the manual of MITE-Hunter, the identified MITE candidates families were first subjected for their coverage evaluation using TARGeT [52]. 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 invert repeats (TIR) and target site duplication (TSD). In total, 15 MITE families were identified. We then assigned these 15 MITE families to different super-families and classes based on sequence homology to a P-MITE database [53]. Two MITE families that showed no homology to any known MITE sequences were excluded from 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 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) [54] (hg-Augustus) and MAKER2 (v.2.28) [55], 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 coding 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 contain repeats occupying 10-50% of their entire gene length, we performed an additional search of the plant refseq database using BLASTX. If the gene matches 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 model were retained for downstream analysis. In total, 35,737 gene models from hg-Augustus predictions were kept. 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 they predicted different gene structure for the same gene. Therefore, when the two pipelines predicted different gene structures on 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 identifies homologous regions using BLAST and then predicts 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. Using all predicted gene models from *N. attenuata*, we further predicted protein coding gene models in *N. obtusifolia*, *N. tomentosiformis* and *N. sylvestris* using a homolog-based approach.

RNA sequencing, data analysis, transcriptome assembly

N. attenuata 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 TURBO™ 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) [56] with parameters "--collapse --trimns --trimqualities 2 --minlength 36". The trimmed reads were then aligned to the *N. attenuata* genome assembly (v 2.0) using TopHat2 (v2.1.0) [57], with maximum and minimum intron size set to 50,000 and 41 base pairs (bp), respectively, estimated from the *N. attenuata* genome annotation. The genes and transcripts were assembled using Cufflinks (v2.2.0) [58] with the *N. attenuata* genome annotation as the reference.

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) [59]. 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.

To study the tissue-specificity of expressed genes/transcripts, we calculated the tissue-specificity score for each gene/transcript using the τ index [60] based on subset of the RNA-seq libraries.

Comparative genomic analysis

We assigned homolog groups (HGs) using a similarity-based method. For this, we used all genes that were predicted from the 11 genomes, listed in Supplementary Table S2. 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 markov cluster algorithm (MLC) [61].

We constructed a phylogenetic tree for all identified HGs using an in-house developed pipeline. In brief, we aligned all coding sequences for each HG using MUSCLE (v.3.8.31) [62], based on translated protein sequences from TranslatorX (v.1.1) [63]. For all aligned sequences, all non-informational sites (gaps in more than 20% of sequences) were removed using trimAL (v1.4) [64]. Then, for each HG, PhyML (v. 20140206) [65] was used to construct the gene tree with the best nucleotide substitution model, estimated based on jModeltest2 (v.2.1.10) [66] with the following parameters: -f -i -g 4 -s 3 -AIC -a. The support for each branch was calculated using the approximate Bayes 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 [67].

Evolution of nicotine biosynthesis and GCC and G-box transcription factor binding sites

Genes previously characterized for implication 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 [3,9], and used FIMO [68] 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 annotation 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 the TE consensus sequences. The number of GCC box and G-box motifs per kilo-base sequences from TE consensus can be as high as 19 and 28, respectively. Permutation tests by randomly shuffling the positions of GCC and G-box 1000 times in *N. attenuata* genome and then compared with TE locations further suggested that these two motifs were significantly enriched in TE regions ($p < 0.001$). These data suggest 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 within the upstream regions of nicotine biosynthesis genes which were derived from TE insertions. 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 [69], 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.

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 Molecular Ecology Department 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.

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.

Figures

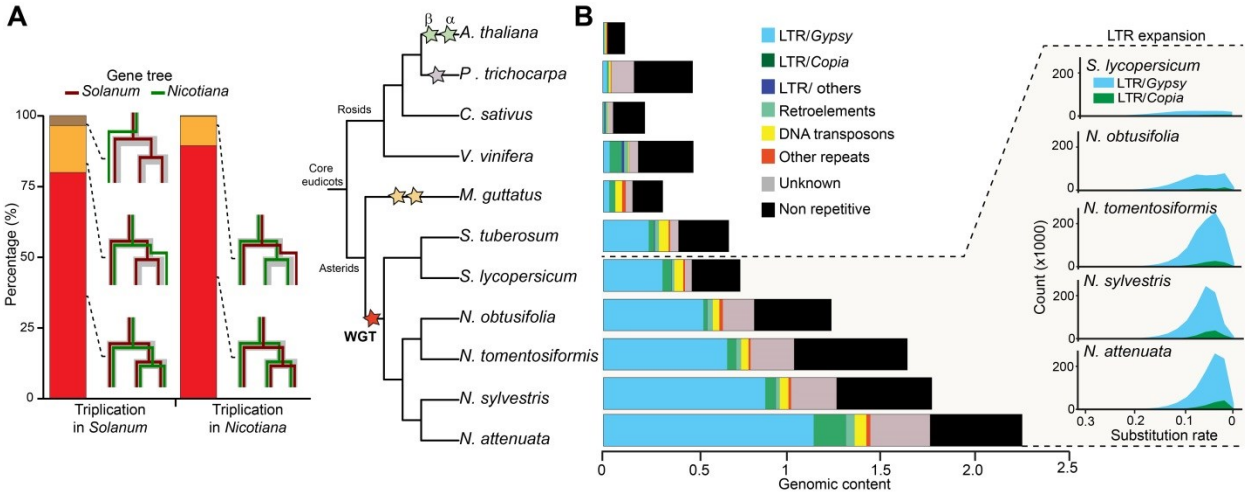


Figure 1. Whole genome triplication (WGT) in *Nicotiana* genomes is shared with other Solanaceae species but the *Gypsy* retrotransposons expansions are *Nicotiana*-specific.

(A) *Nicotiana* genomes share the WGT with other Solanaceae species. Left panel depicts the shared WGT event between *Nicotiana* and *Solanum* as revealed by phylogenetic tree structure of triplicated gene families in *Nicotiana* and *Solanum*. Red and yellow bars represent the percentage of triplication 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.

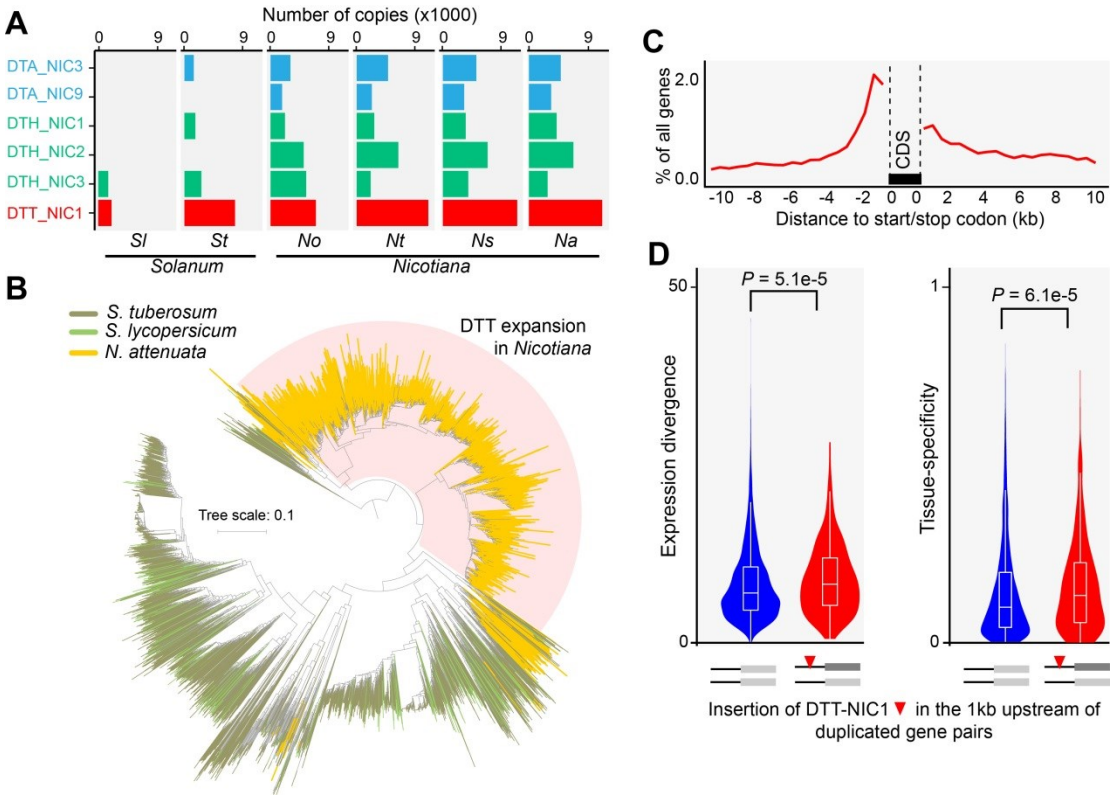


Figure 2. Expansion of transposable elements of the family DTT-NIC1 increased genome-wide gene expression divergence among duplicated gene pairs in *Nicotiana*.

(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 5 main MITE transposable elements (TEs). MITE families are visualized by different colors: light blue, DTA (*h*AT); 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. obtusifolia*; *Ns*, *N. sylvestris*; *Nt*, *N. tomentosiformis*. *Solanum* species: *Sl*, *Solanum lycopersicum*; *St*, *Solanum 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. Divergence of expression and tissue specificity were calculated from expression data of 20 different RNA-seq libraries from different tissues or same tissue with different treatments. 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.

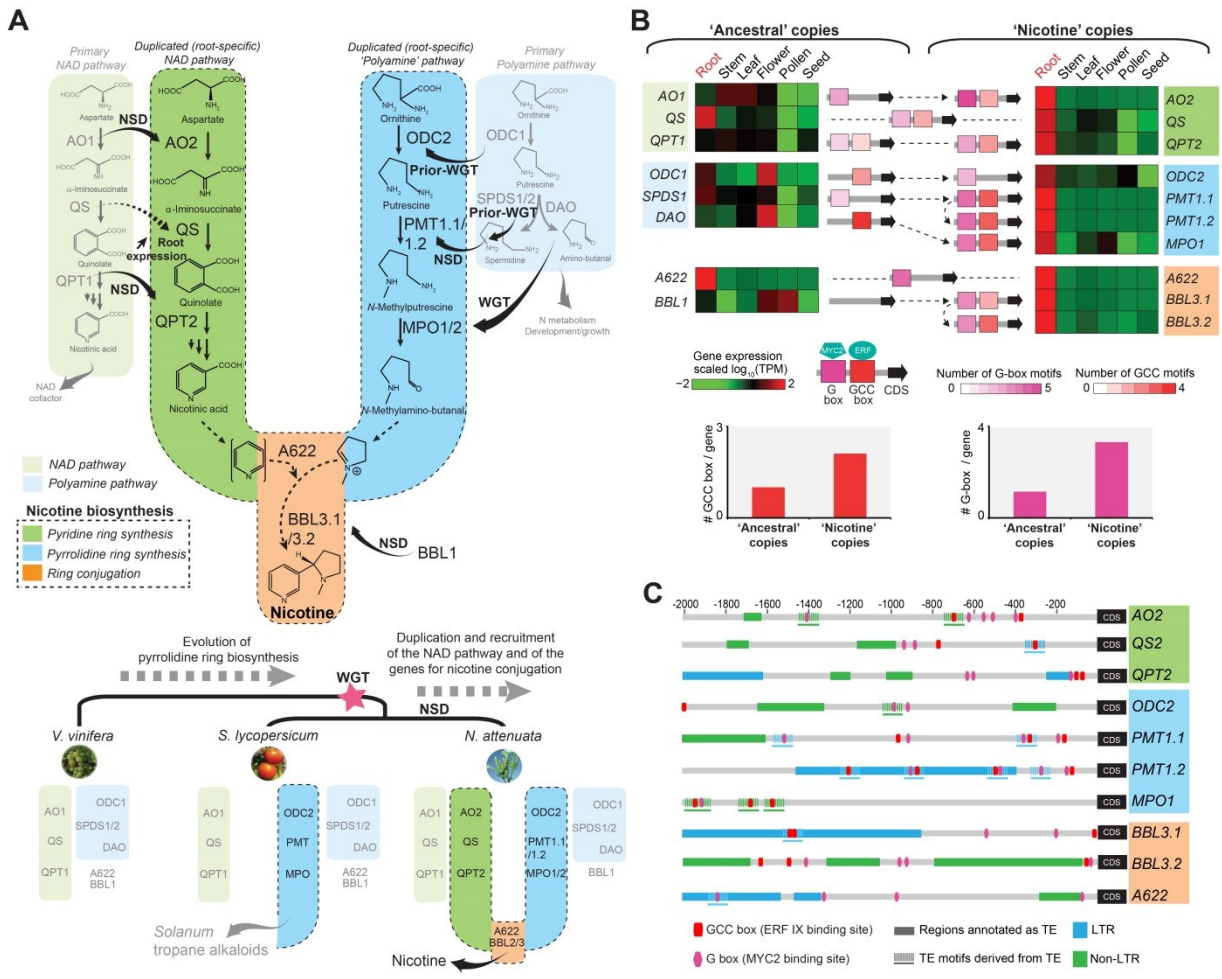


Figure 3. Prolific nicotine production evolved from the duplication of two primary metabolic pathways and its coordinated transcriptional regulation was likely facilitated by transposon-derived transcription factor binding site insertions.

(A) Nicotine biosynthesis genes originate 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: phylogenomics 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 depict the scaled expression of nicotine biosynthetic genes and their ancestral copies across six distinct tissues. Red and green signify high and low expression, respectively. 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 derived from TE. Each row depicts the motif and TE annotation of the 2 kb upstream region of an individual nicotine biosynthesis gene. GCC and G-box motifs were predicted using FIMO with E-value less than 1e-3 and 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.

1. Mithofer, A. & Boland, W. Plant defense against herbivores: chemical aspects. *Annu. Rev. Plant Biol.* **63**, 431-450 (2012).
2. Swain, T. Secondary compounds as protective agents. *Annu Rev Plant Phys* **28**, 479-501 (1977).
3. Shoji, T. *et al.* Clustered transcription factor genes regulate nicotine biosynthesis in tobacco. *Plant Cell* **22**, 3390-3409 (2010).
4. Kessler, A. *et al.* Silencing the jasmonate cascade: induced plant defenses and insect populations. *Science* **305**, 665-8 (2004).
5. Kessler, D. *et al.* Field experiments with transformed plants reveal the sense of floral scents. *Science* **321**, 1200-2 (2008).
6. Steppuhn, A. *et al.* Nicotine's defensive function in nature. *PLoS Biol.* **2**, E217 (2004).
7. Kessler, D. *et al.* Unpredictability of nectar nicotine promotes outcrossing by hummingbirds in *Nicotiana attenuata*. *Plant J.* (2012).
8. Adler, L.S. *et al.* Reliance on pollinators predicts defensive chemistry across tobacco species. *Ecol Lett* **15**, 1140-1148 (2012).
9. Shoji, T. & Hashimoto, T. Tobacco MYC2 regulates jasmonate-inducible nicotine biosynthesis genes directly and by way of the NIC2-locus ERF genes. *Plant Cell Physiol.* **52**, 1117-1130 (2011).
10. Baldwin, I.T. *et al.* Allocation of N-15 from nitrate to nicotine - production and turnover of a damage-induced mobile defense. *Ecology* **75**, 1703-1713 (1994).
11. Baldwin, I.T. & Callahan, P. Autotoxicity and chemical defense: nicotine accumulation and carbon gain in solanaceous plants. *Oecologia* **94**, 534-541 (1993).
12. Cowley, M. & Oakey, R.J. Transposable elements re-wire and fine-tune the transcriptome. *PloS Genetics* **9**(2013).
13. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49-61 (2013).
14. Chuong, E.B. *et al.* Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083-1087 (2016).
15. Hollister, J.D. & Gaut, B.S. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* **19**, 1419-1428 (2009).
16. Hollister, J.D. *et al.* Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 2322-2327 (2011).
17. Feschotte, C. *et al.* Plant transposable elements: Where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329-341 (2002).
18. Parra, G. *et al.* CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. *Bioinformatics* **23**, 1061-1067 (2007).
19. Sierro, N. *et al.* Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol* **14**(2013).
20. Rausher, M.D. & Huang, J. Prolonged adaptive evolution of a defensive gene in the Solanaceae. *Mol. Biol. Evol.* **33**, 143-151 (2016).
21. Gonzales-Vigil, E. *et al.* Adaptive evolution of threonine deaminase in plant defense against insect herbivores. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5897-5902 (2011).
22. Kang, J.H. *et al.* Silencing threonine deaminase and JAR4 in *Nicotiana attenuata* impairs jasmonic acid-isoleucine-mediated defenses against *Manduca sexta*. *Plant Cell* **18**, 3303-20 (2006).
23. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836-846 (2005).
24. Grandbastien, M. *et al.* Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet. Genome Res.* **110**, 229-241 (2005).

25. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270-278 (2014).
26. Kuang, H.H. *et al.* Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. *Genome Res* **19**, 42-56 (2009).
27. Naito, K. *et al.* Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461**, 1130-U232 (2009).
28. Blanc, G. & Wolfe, K.H. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**, 1679-1691 (2004).
29. Flagel, L.E. & Wendel, J.F. Gene duplication and evolutionary novelty in plants. *New Phytol.* **183**, 557-564 (2009).
30. Kajikawa, M. *et al.* Vacuole-localized berberine bridge enzyme-like proteins are required for a late step of nicotine biosynthesis in tobacco. *Plant Physiol.* **155**, 2010-2022 (2011).
31. Kajikawa, M. *et al.* A PIP-family protein is required for biosynthesis of tobacco alkaloids. *Plant Mol. Biol.* **69**, 287-298 (2009).
32. Minguet, E.G. *et al.* Evolutionary diversification in polyamine biosynthesis. *Mol. Biol. Evol.* **25**, 2119-2128 (2008).
33. Shoji, T. *et al.* Jasmonate-induced nicotine formation in tobacco is mediated by tobacco COI1 and JAZ genes. *Plant Cell Physiol.* **49**, 1003-1012 (2008).
34. Machado, R.A. *et al.* Benefits of jasmonate-dependent defenses against vertebrate herbivores in nature. *Elife* **5**(2016).
35. Krugel, T. *et al.* Agrobacterium-mediated transformation of *Nicotiana attenuata*, a model ecological expression system. *Chemoecology* **12**, 177-183 (2002).
36. Bubner, B. *et al.* Two-fold differences are the detection limit for determining transgene copy numbers in plants by real-time PCR. *BMC Biotechnol.* **4**(2004).
37. Anssour, S. *et al.* Phenotypic, genetic and genomic consequences of natural and synthetic polyploidization of *Nicotiana attenuata* and *Nicotiana obtusifolia*. *Ann. Bot.* **103**, 1207-1217 (2009).
38. VanBuren, R. *et al.* Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508-11 (2015).
39. Kallenbach, M. *et al.* Empoasca leafhoppers attack wild tobacco plants in a jasmonate-dependent manner and identify jasmonate mutants in natural populations. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1548-E1557 (2012).
40. Heiling, S. *et al.* Jasmonate and ppHsystemin regulate key malonylation steps in the biosynthesis of 17-hydroxygeranyllinalool diterpene glycosides, an abundant and effective direct defense against herbivores in *Nicotiana attenuata*. *Plant Cell* **22**, 273-92 (2010).
41. Leitch, I.J. *et al.* The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann. Bot.* **101**, 805-814 (2008).
42. Boetzer, M. *et al.* Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579 (2011).
43. Luo, R.B. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**(2012).
44. English, A.C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**(2012).
45. Walker, B.J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**(2014).
46. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**(2012).

47. Peng, Y. *et al.* IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428 (2012).
48. Abrusan, G. *et al.* TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329-1330 (2009).
49. Sato, S. *et al.* The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).
50. Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-U94 (2011).
51. Han, Y.J. & Wessler, S.R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* **38**(2010).
52. Han, Y.J. *et al.* TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences. *Nucleic Acids Res* **37**(2009).
53. Chen, J.J. *et al.* P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res* **42**, D1176-D1181 (2014).
54. Stanke, M. *et al.* AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* **32**, W309-W312 (2004).
55. Cantarel, B.L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188-196 (2008).
56. Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes* **5**, 337 (2012).
57. Trapnell, C. *et al.* TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
58. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-78 (2012).
59. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
60. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650-9 (2005).
61. Enright, A.J. *et al.* An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).
62. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
63. Abascal, F. *et al.* TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* **38**, W7-W13 (2010).
64. Capella-Gutierrez, S. *et al.* trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
65. Guindon, S. *et al.* Estimating maximum likelihood phylogenies with PhyML. *Bioinformatics for DNA Sequence Analysis* **537**, 113-137 (2009).
66. Darriba, D. *et al.* jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772-772 (2012).
67. Page, R.D.M. & Charleston, M.A. From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem. *Mol Phylogenet Evol* **7**, 231-240 (1997).
68. Grant, C.E. *et al.* FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).
69. Noe, L. & Kucherov, G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* **33**, W540-W543 (2005).

Supplementary Information

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 of 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 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). **Supplementary Table S1** 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 were 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 (**Supplementary Figure S1**). Molecules were *de novo* assembled as previously described [2]. The optical map finally assembled consists of 2.2 Gb with a 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 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 mapc

A high-density linkage map was constructed using the genotyping by sequencing (GBS) method on 256 individuals from an advanced inter-cross recombinant inbred line population (AI-RIL). The establishment of the AI-RIL and detailed procedures on DNA isolation and sequencing will be described in detail elsewhere. 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 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 MQ0 >= 4 && ((MQ0 / (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 13,076 (38.9%) of total predicted genes were anchored to 12 pseudomolecules (**Supplementary Figure 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 **Supplementary Figure 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 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 (**Supplementary Table S2**). Fourth, re-sequencing of ~2kb the up-stream 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 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 **Supplementary Table S3**. Abundance of MITEs are shown in **Supplementary Dataset S1**. 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 identify the consequences of DTT-NIC1 insertions on genome-wide gene expression divergences and the resulting effects on herbivory-induced defense signaling, two different datasets were used. For the analysis of herbivory-induced defense signaling, we mined a time-course dataset based on microarray[11] of the *Manduca sexta* oral secretion (OS)-induced transcriptomic changes in *N. attenuata* leaves and compared the likelihood of being significantly induced by OS between genes with DTT-NIC1 insertion in their 1kb upstream region and genes without. The results showed that DTT-NIC1 insertions significantly increased the likelihood of genes of being induced at 1, 13, and 21 h after elicitation (**Supplementary Figure S4**). Additional analyses of 60 herbivore associated elicitors induced leaf transcriptomic responses among six closely related *Nicotiana* species further support the conclusion that DTT-NIC1 insertions contributed to the recruitment of genes to herbivore induced early defenses signaling in *Nicotiana* (Zhou et al. submitted). To analyze 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 analysed the gene duplication history for all of *N. attenuata* genes based on the constructed phylogenetic trees of each homolog group; 2) gene duplications that have generated a reciprocal most recently duplicated paralog pairs were identified; 3) only the gene pairs when the tree branch that provide support for estimating duplication events has an approximate Bayesian support greater than 0.9 were kept for downstream analysis; 4) the gene pairs were then assigned into two groups, one group as the gene pairs of which one of the two copies had at least one DTT-NIC1 insertion (within 1kb upstream of 5' region) and the other group as the gene pairs where neither copy had DTT-NIC1 insertions. Gene pairs that both had DTT-NIC1 insertions were excluded from downstream analysis; 5) the gene expression divergence at both expression and tissue specificity levels were then compared between the two identified groups.

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 reads mapping 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 "mapable 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 mapable reads. Conserved miRNAs were identified by comparing the *N. attenuata* mapable 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 S2**. 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.4 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 (**Supplementary Table S4**). Expression profile of all predicted gene models is provided in **Supplementary dataset S3.** 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 S4**). Among all sequenced tissues, roots and pollen tubes expressed the highest and lowest number of both genes and transcripts (**Supplementary Figure S5**), 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 S4**), 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 τ index >= 0.95. Among all sequenced tissues, roots have the highest number of tissue-specific genes and transcripts (**Supplementary Figure S6**). 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 JuncBASE (v0.6) [14] based on the splicing junctions (SJs) 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 (**Supplementary Figure S7**).

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 (**Supplementary Figure S8, A and B**).

For *M. sexta* herbivory-induced TE expression in leaves, we used a previously published time-course microarray dataset[11]. 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 probes and repeats were identified using bedtools [17] based on repeatmakser 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[18]. 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 S5**. Similar to protein coding genes, most of TEs showed highest induction at 1 h after elicitation (**Supplementary Figure S9**), 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[19]. 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 code 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 [20], 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 S6**.

2.7 Functional annotation of protein domains

To identify the functional domains for protein-coding genes, INTERPROSCAN (v. 5.16-55.0)[21] 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 **Supplementary Figures S10 and S11**. The top 20 most frequent domains accounts 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[22], 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-Rodríguez et al.,[23] using 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 S7**). Among all transcription factors, MYB, AP2 and bHLH were the three largest families.

3. Evolutionary genomics analyses

3.1 Gene homolog group identification

In total, we identified 23,340 HGs (with at least two homolog sequences) from the 11 plant genomes (**Supplementary Table S5**). 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 (**Supplementary Figure S12**).

3.2 Detection of gene duplication events in each HG

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[24]. 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 (**Supplementary Figure S13, and Supplementary dataset S8**). Furthermore, a large number of gene duplication events were found on the branch of Solanaceae, indicating the shared WGT [22] between the *Nicotiana* and *Solanum* genera (**Supplementary Figure S13**). One example 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 (**Supplementary Figure S14**).

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* [22], 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) [25]. To estimate the divergence time, the formula T=Ks/ (2*mutational rate) was used, where the mutational rate was set to $7.1*10^{-9}$ substitutions/site/million years (MYA)[26]. 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 (**Supplementary Figure S15**). 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)[22] 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 (**Supplementary Figure S16**). In addition, we also compared the identified triplication events in *Solanum* and *Nicotiana* with those of genes in *Mimulus*. Our results showed that a majority of the triplication events (98.7% and 98.4% in *Solanum* and *Nicotiana,* respectively) were not shared with *Mimulus*, indicating that gene duplication events in *Mimulus* are independent of the WGT found in Solanaceae.

3.4 Estimation of species divergence times

We calculated the divergence times of four Solanacea 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)[27] based on translated protein sequences from TranslatorX (v.1.1)[28]. These alignments were concatenated to one super-alignment and all ambiguously aligned regions were removed using trimAL (v. 1.4) [29] 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)[30]. 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[31]. 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[32]; 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[33], 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 expanding significantly 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[34]. 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 (**Supplementary Figure S17**). 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[35]. 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 (**Supplementary Figure S18**). 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[36,37]. 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 S9**). Among these significantly expanded gene families, a *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 (**Supplementary Figure S19**). 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 permease*s (**Supplementary Figure S20**). Genes from this family are known as plasma membrane-localized transporters[38]. More specifically, in tobacco, one member of this family, *Nicotine Uptake Permease1* (NUP1), has demonstrated functions in regulating nicotine localization via its transporter activity[38,39]. Furthermore, *NUP1* has also been shown to act as a transcriptional regulator of the key transcription factor ERF189 in the nicotine biosynthesis pathway[39], 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 (**Supplementary Figure S20**), 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 [40]. 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 (**Supplementary Table S6**). Gene evolutionary history were inferred using the combination of phylogenetic and synteny analysis when possible (**Supplementary Figure S21-S28**).

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 [41,42]. 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 (**Supplementary 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 detected between Sanger sequencing and the assembled genome.

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

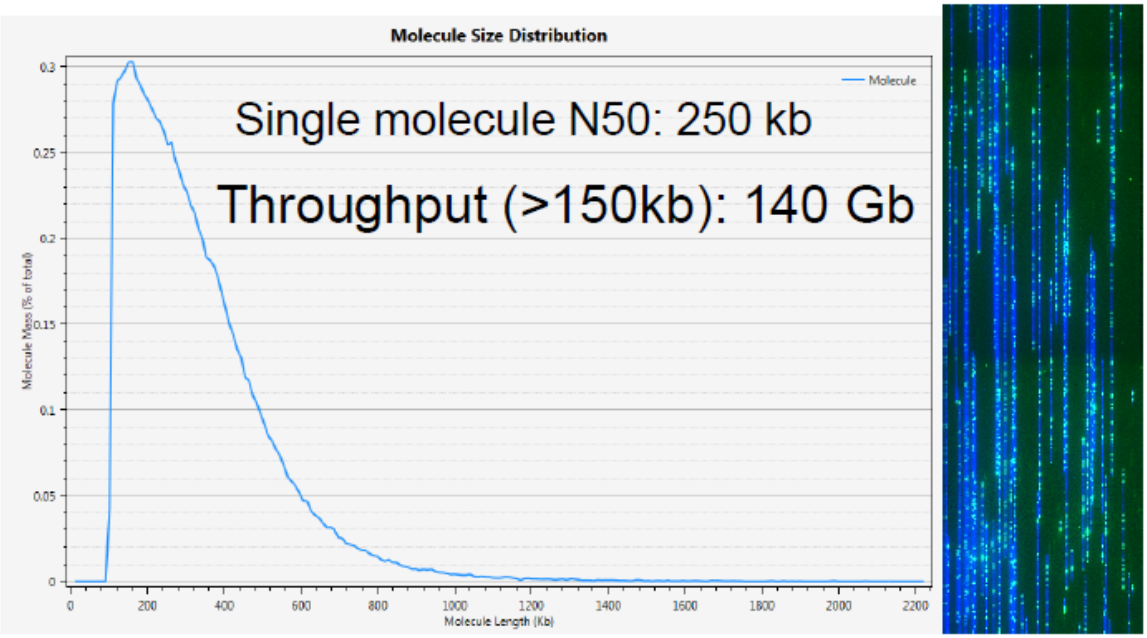
Insertions of TE are known to introduce putative target sites of miRNAs. To examine whether the observed TE insertions within 2kb upstream region of nicotine biosynthesis genes and of their ancestral copies had introduced candidate miRNA targeting sites, we performed *in silico* miRNA target site predictions using the method described in Pandey et al [43]. 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 mis-matches 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 (**Supplementary Table S8**). Among all nicotine biosynthesis genes, five predicted miRNA target sites that were detected from 4 genes (*BBL3.2, PMT1.2, ODC2 and QS*) are overlapped with TE insertions.

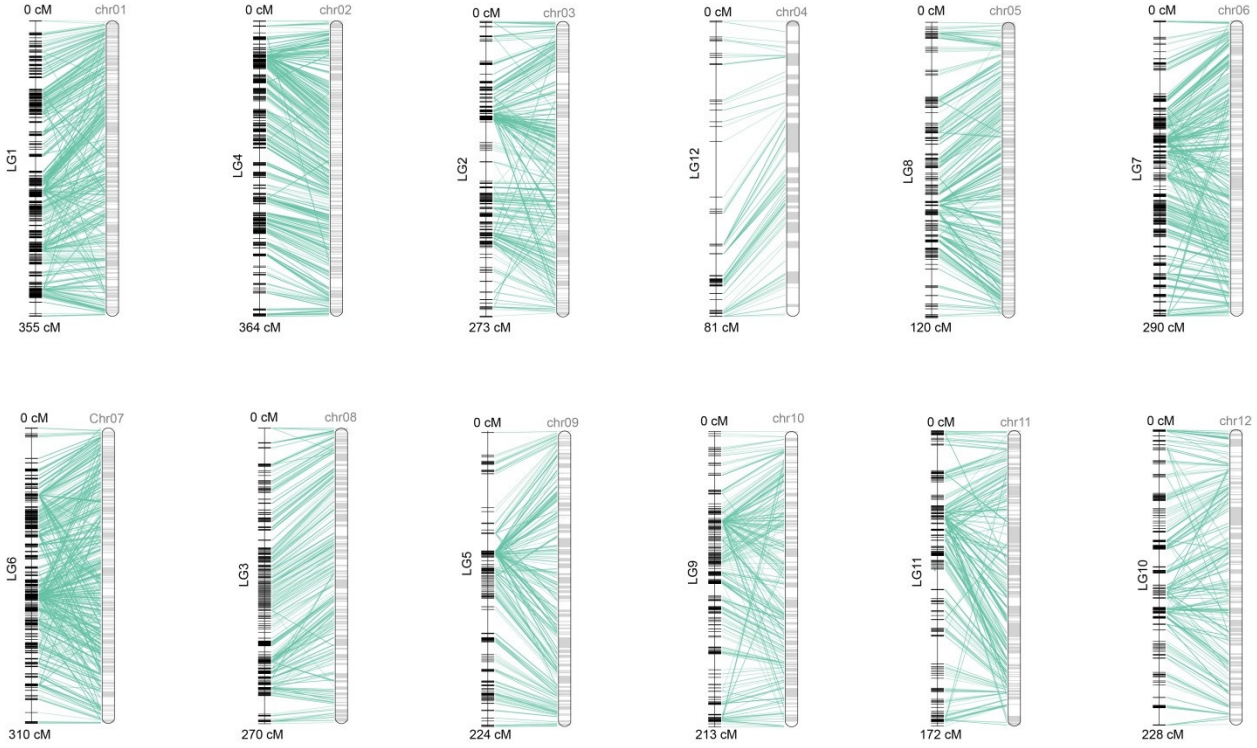
4.4 Data access

To provide free access to our genome data, we established a webserver (*Nicotiana attenuata* Data Hub: http://nadh.ice.mpg.de/), 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 public available via SGN and CoGe. Including or excluding these gene models does not affect any conclusion of the study.

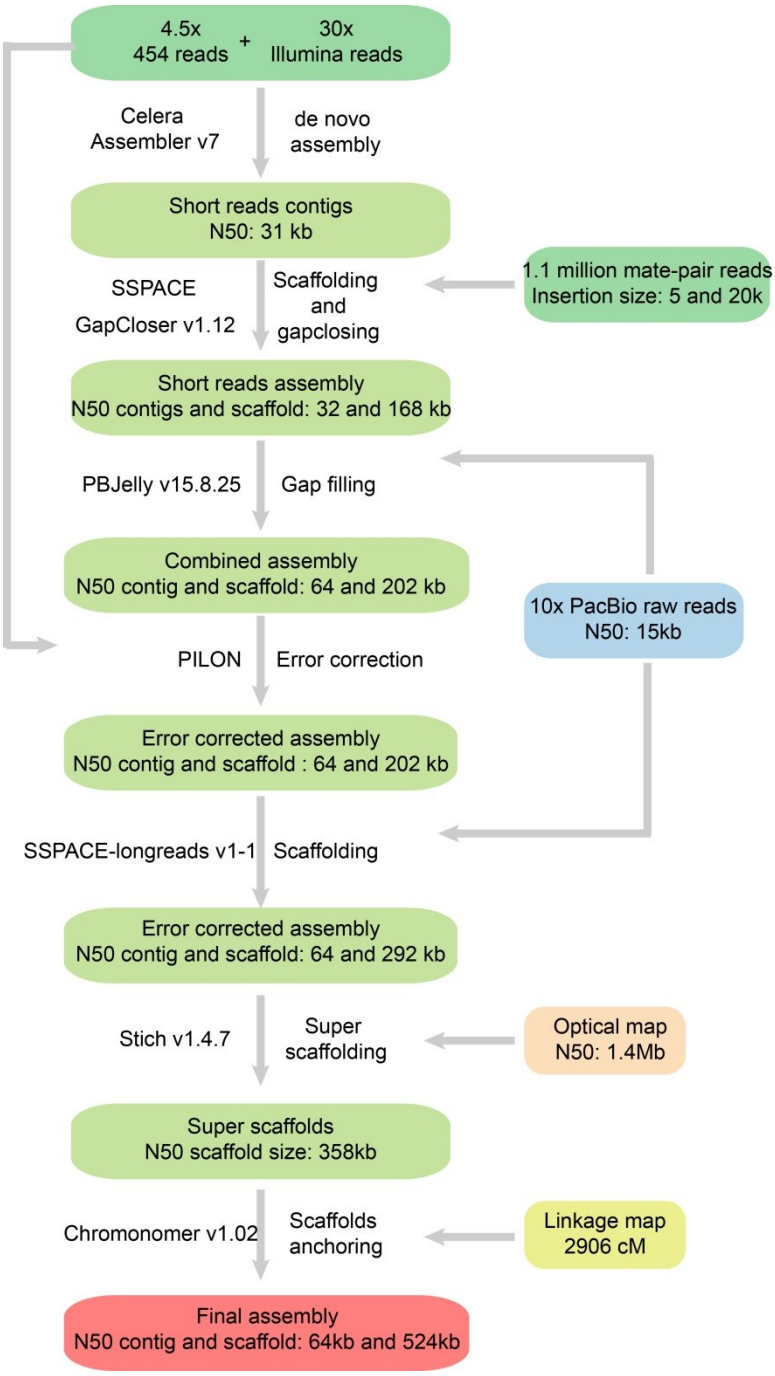
5. Supplemental Figures.



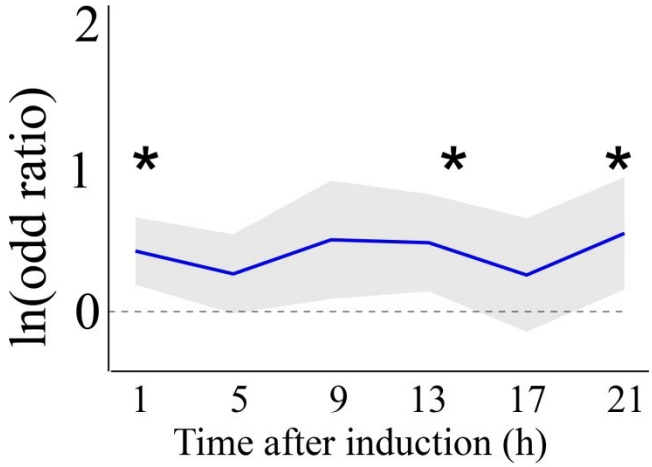
Supplementary Figure S1. Distribution of Bio-Nano molecule length. Left panel indicates the size distribution of the BioNano molecule length, right panel shows an example of the BioNano output image.



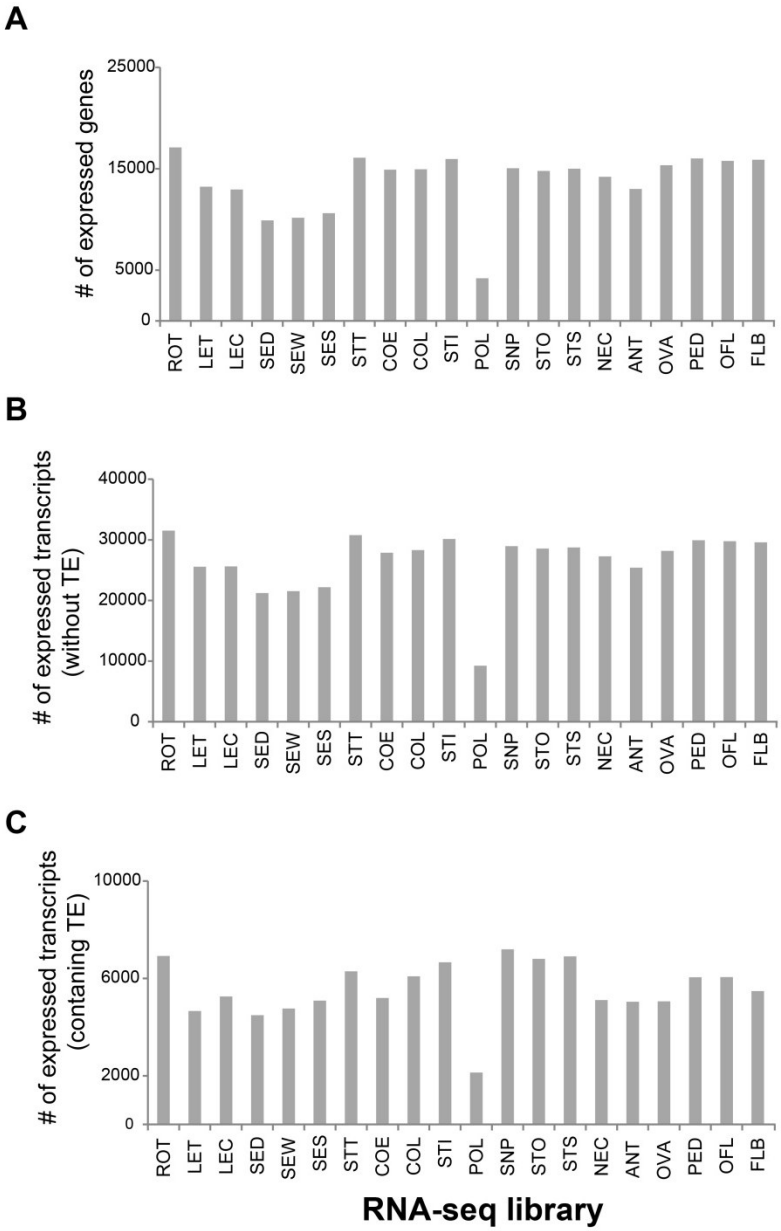
Supplementary Figure S2. Anchoring of scaffolds to the linkage map. In total, 12 linkage groups were constructed. Linkage map and anchored pseudo-chromosomes are shown on left and right sides, respectively.



Supplementary Figure 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.

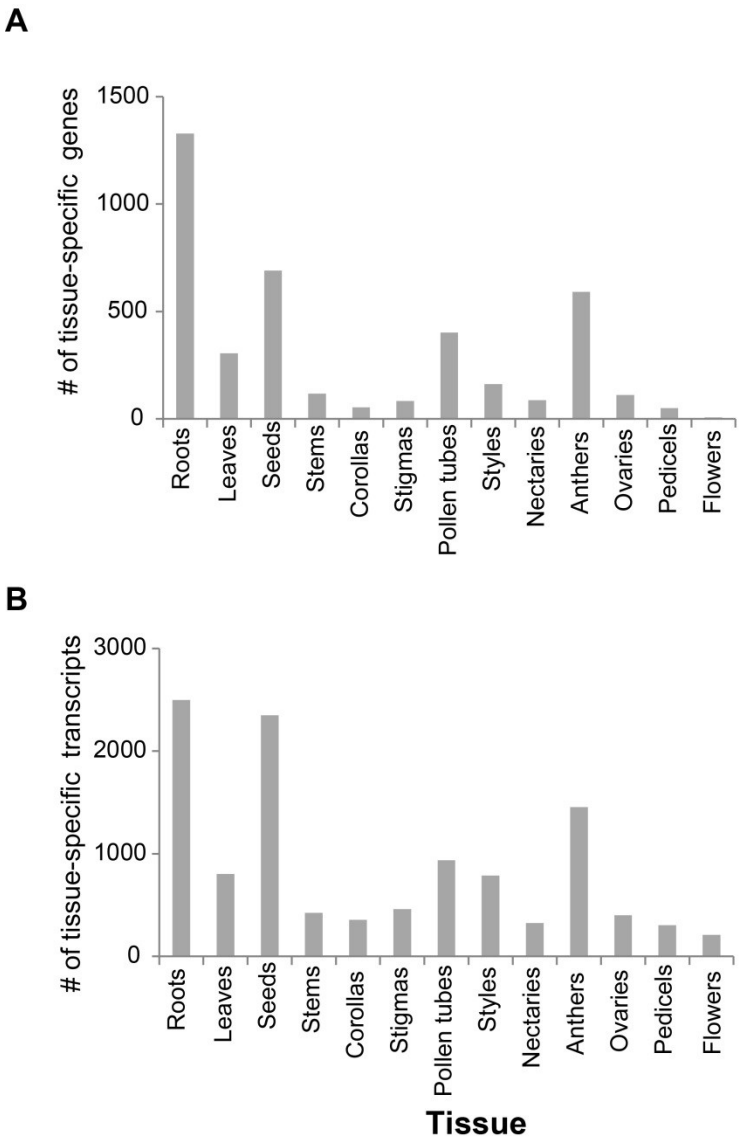


Supplementary Figure S4. Enrichment of DTT-NIC1 in the 1kb upstream region of *Manduca sexta* oral secretions (OS) up-regulated genes. The Y-axis indicates the natural log odd ratio of *M. sexta*-induced genes. Odd ratios were calculated using this formula: Odd = (p1/(1 – p1))/(p2/(1 – p2)), where p1 and p2 are the probability of genes induced by OS among genes that harbor DTT-NICl insertions within their 1kb upstream region and probability of genes induced by OS among all genes. Values above 0 indicate positive enrichment. Ribbons indicate the 95% confidence intervals. Asterisks indicate statistical significance ($P < 0.01$, Fisher's exact test).

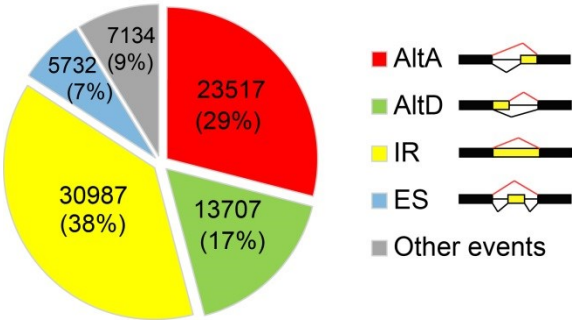


Supplementary Figure S5. Distribution of expressed genes and transcripts among different tissue-specific RNA-seq libraries. A) number of expressed genes; **B)** number of expressed total transcripts; **C)** number of expressed transcripts that contain transposon sequences. The abbreviations of sample 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

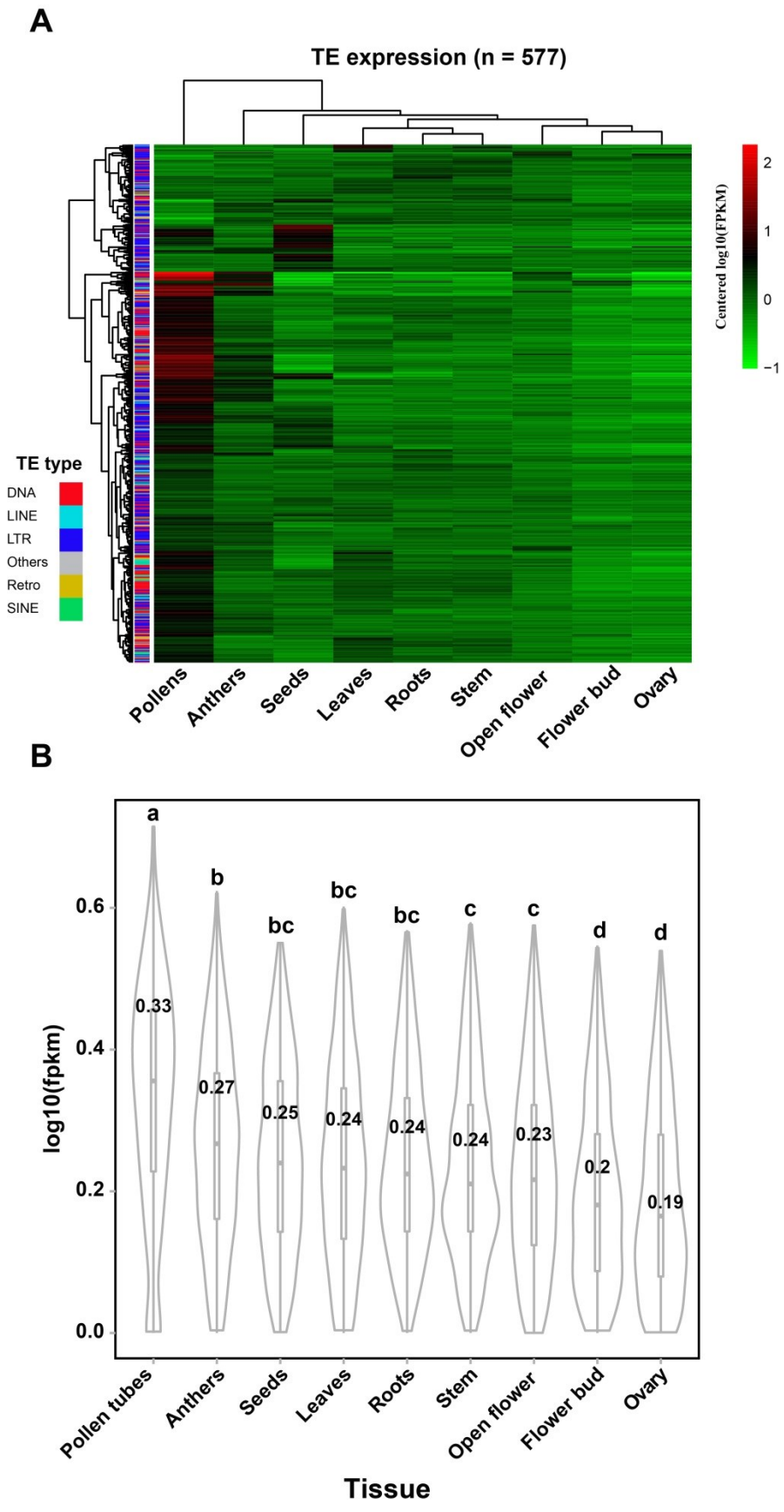
stage; COL: corollas at late developmental stage; STI: styles; POL: pollen tubes grown in pollen germination media; SNP: stigmas not pollinated; STO: stigmas outcrossed; STS: stigmas self-pollinated; NEC: nectaries; ANT: anthers; OVA: ovaries; PED: pedicels; OFL: open flower; FLB: flower bud.



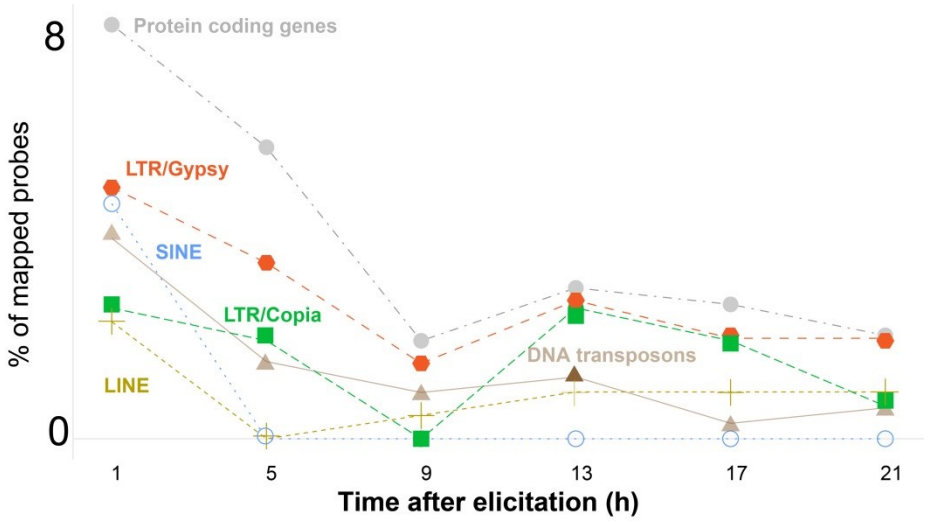
Supplementary Figure S6. Distribution of tissue-specific expressed genes and transcripts. A) number of tissue specifically expressed genes; **B)** number of tissue specifically expressed transcripts.



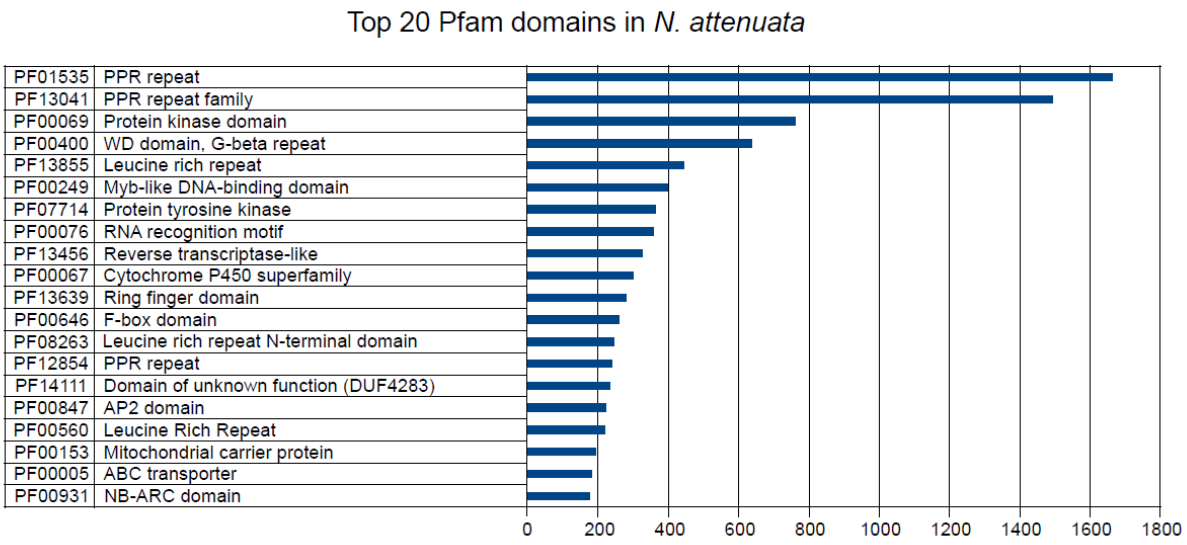
Supplementary Figure S7. Distribution of different alternative splicing events in *N. attenuata*. AltA refers to alternative acceptor site; AltD refers to alternative donor site; IR refers to intron retention and ES refers to exon skipping.


A
TE expression (n = 577)
Centered log10(FPKM)
2
1
0
−1
TE type
DNA
LINE
LTR
Others
Retro
SINE
Pollens
Anthers
Seeds
Leaves
Roots
Stem
Open flower
Flower bud
Ovary
B
log10(fpkm)
0.0
0.2
0.4
0.6
a
b
bc
bc
bc
c
c
d
d
0.33
0.27
0.25
0.24
0.24
0.24
0.23
0.2
0.19
Pollen tubes
Anthers
Seeds
Leaves
Roots
Stem
Open flower
Flower bud
Ovary
Tissue

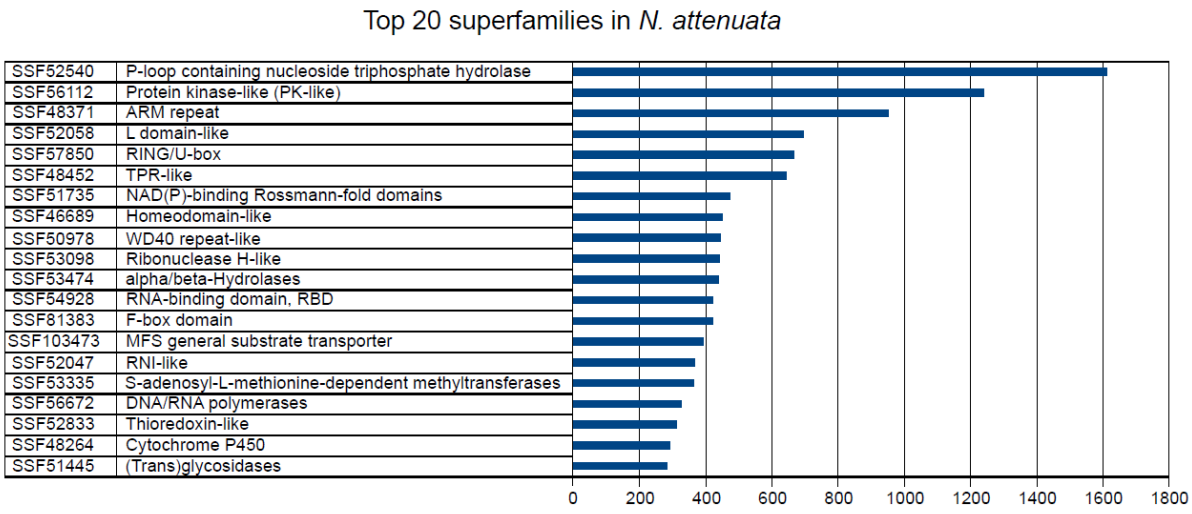
Supplementary Figure S8. Expression of transposable elements among nine tissues. A) heatmap depicting the relative expression of 577 TE families that showed expression in at least one tissue (FPKM>1). Classification of the different TEs is shown with different colors on left side. **B)** boxplots showing 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.



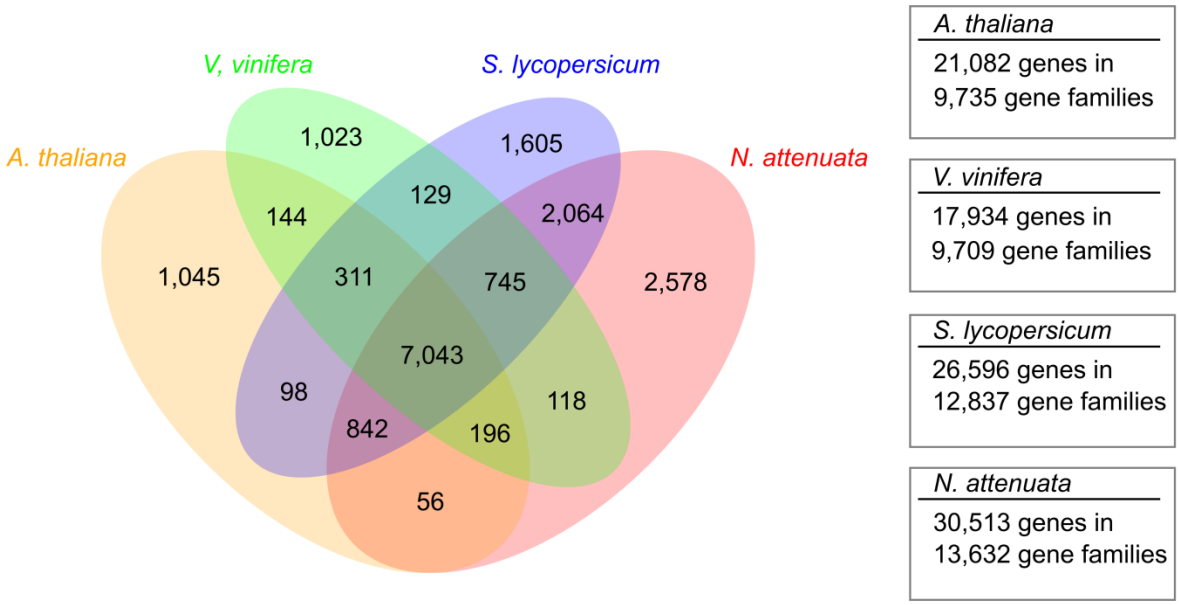
Supplementary Figure S9. Percentage of TEs induced by *M. sexta* oral secretions in *N. attenuata* leaves within a 21 h time course. Each color indicates different classes of TE families.



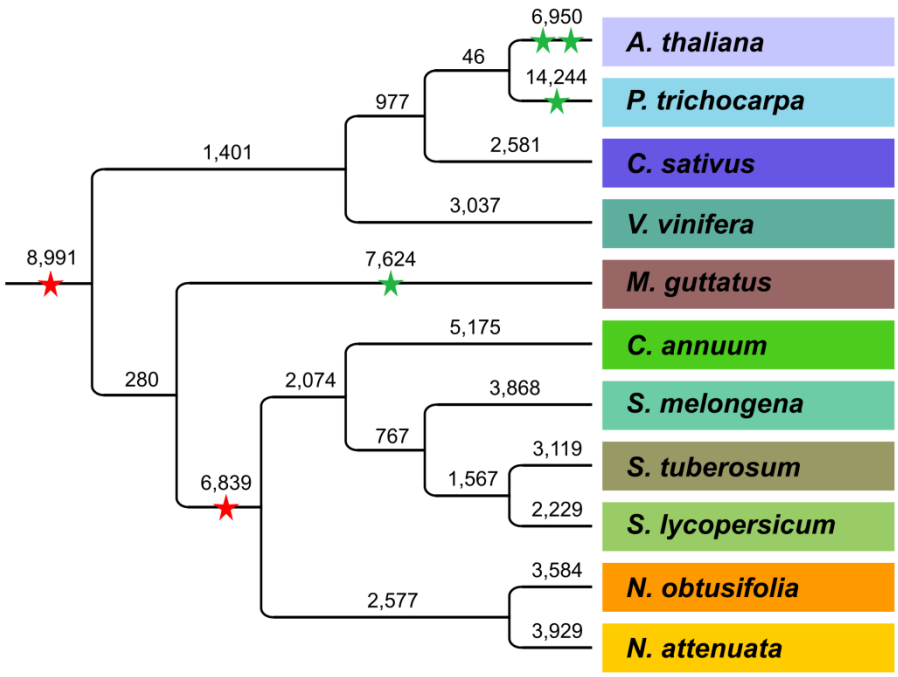
Supplementary Figure S10. The most abundant 20 pfam domains in *N. attenuata*.



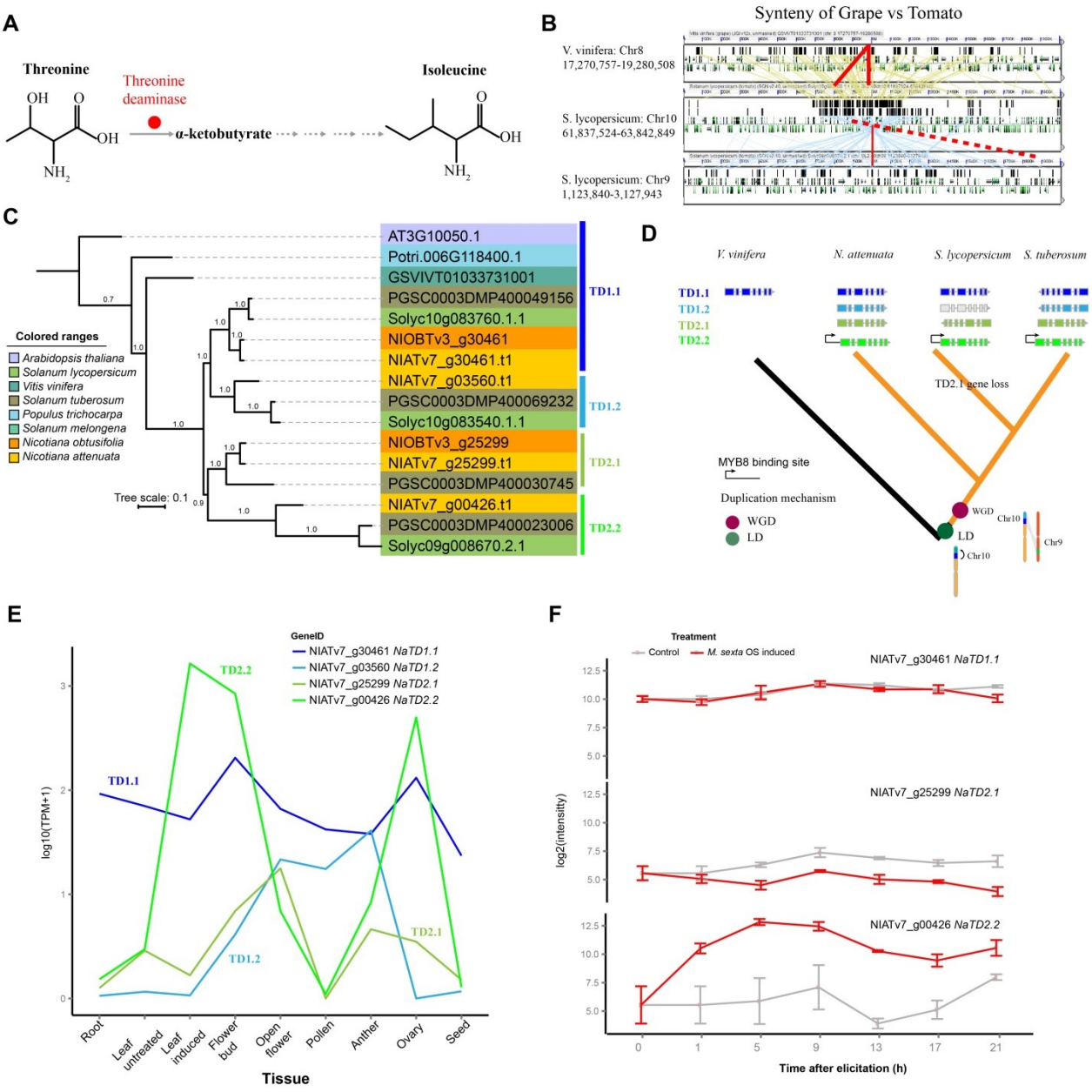
Supplementary Figure S11. Top 20 superfamily domains in *N. attenuata*.



Supplementary Figure S12. Distribution of gene families among four plant species.

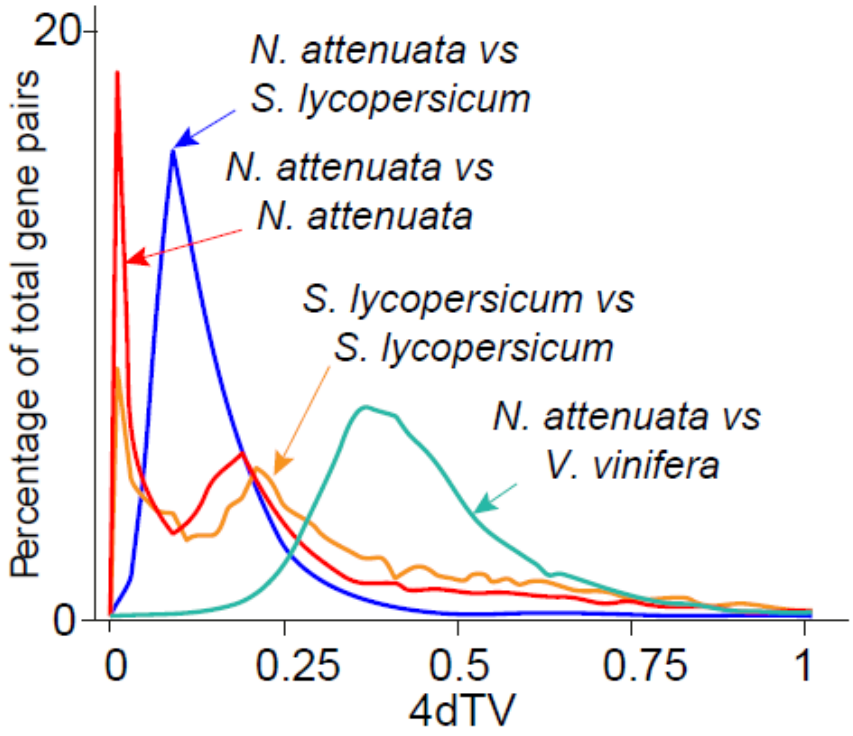


Supplementary Figure S13. Gene family size evolution among 11 plant genomes. The number of all duplication events (above branch nodes) from 11 different plant species. The duplication events were estimated from phylogenetic trees constructed from homolog groups. Only branches that have approximated Bayes branch support values greater than 0.9 are presented. Known whole genome triplication (red stars) and duplication (green stars) events are shown.

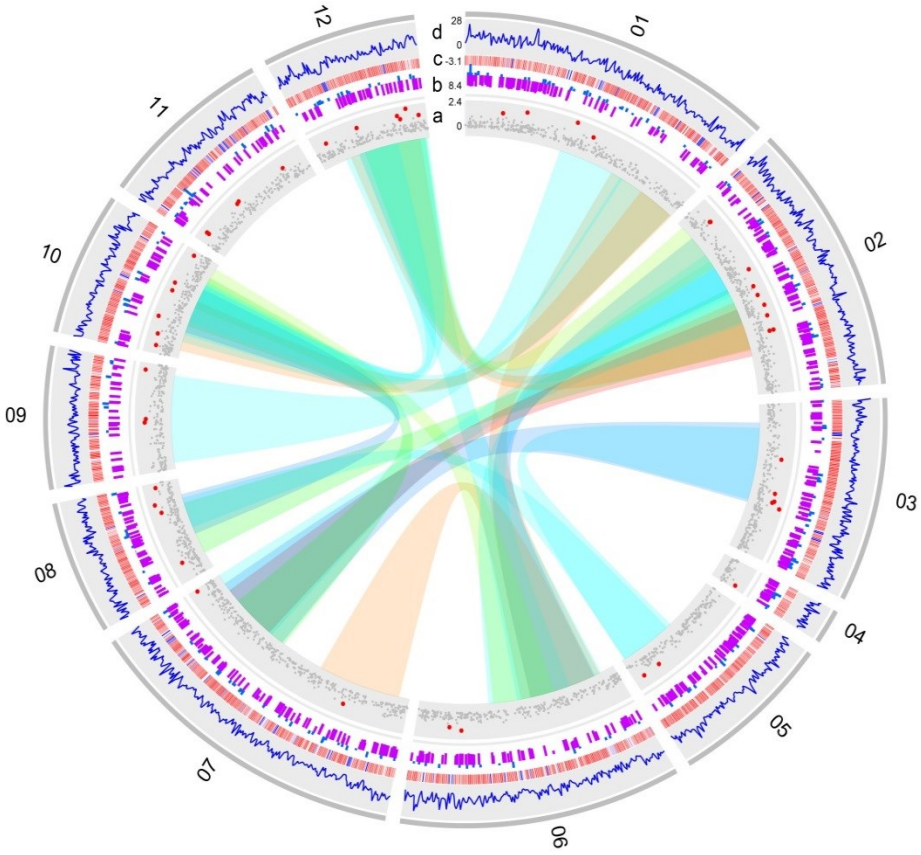


Supplementary Figure S14. Evolution of *threonine deaminase* (TD). A) molecular function of TD, which converts threonine to α- ketobutyrate, the substrate for the biosynthesis of isoleucine. **B**) syntenic information of TD in grape and tomato genomes. Upper panel shows the syntenic region between grape and tomato chromosome 9. Two copies of TD were found on tomato chromosome 10 (between 61.8Mb-63.8Mb), and one copy on tomato chromosome 9, which was reverse duplicated from chromosome 10. **C**

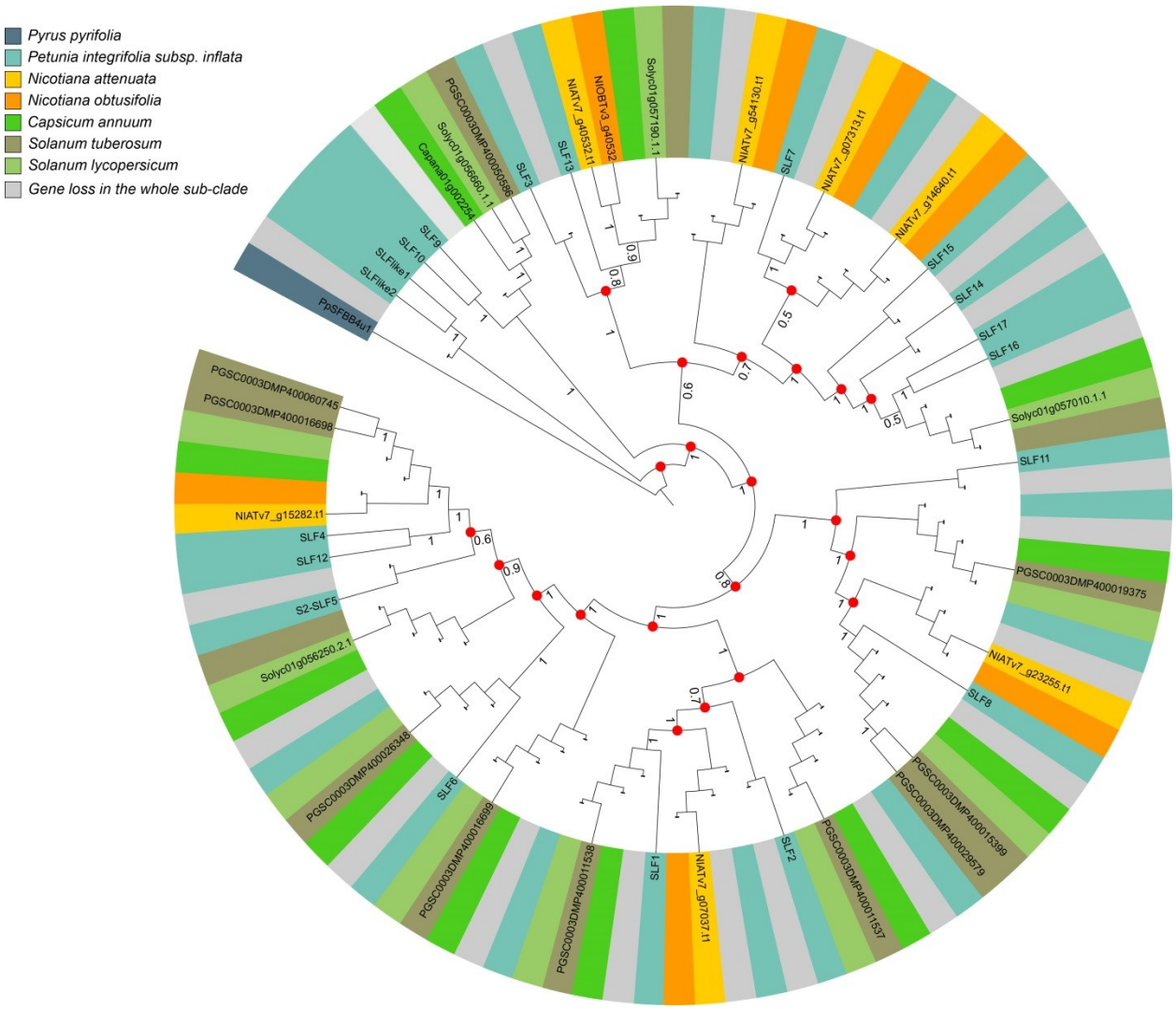
and D) Phylogenetic tree of the TD family (C) and a simplified model (D) show the evolutionary history of TD. Numbers on each branch indicate the approximate Bayes branch supports. Local duplication and whole genome duplication events are shown as purple and green circles, respectively. **E** and **F)** 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).



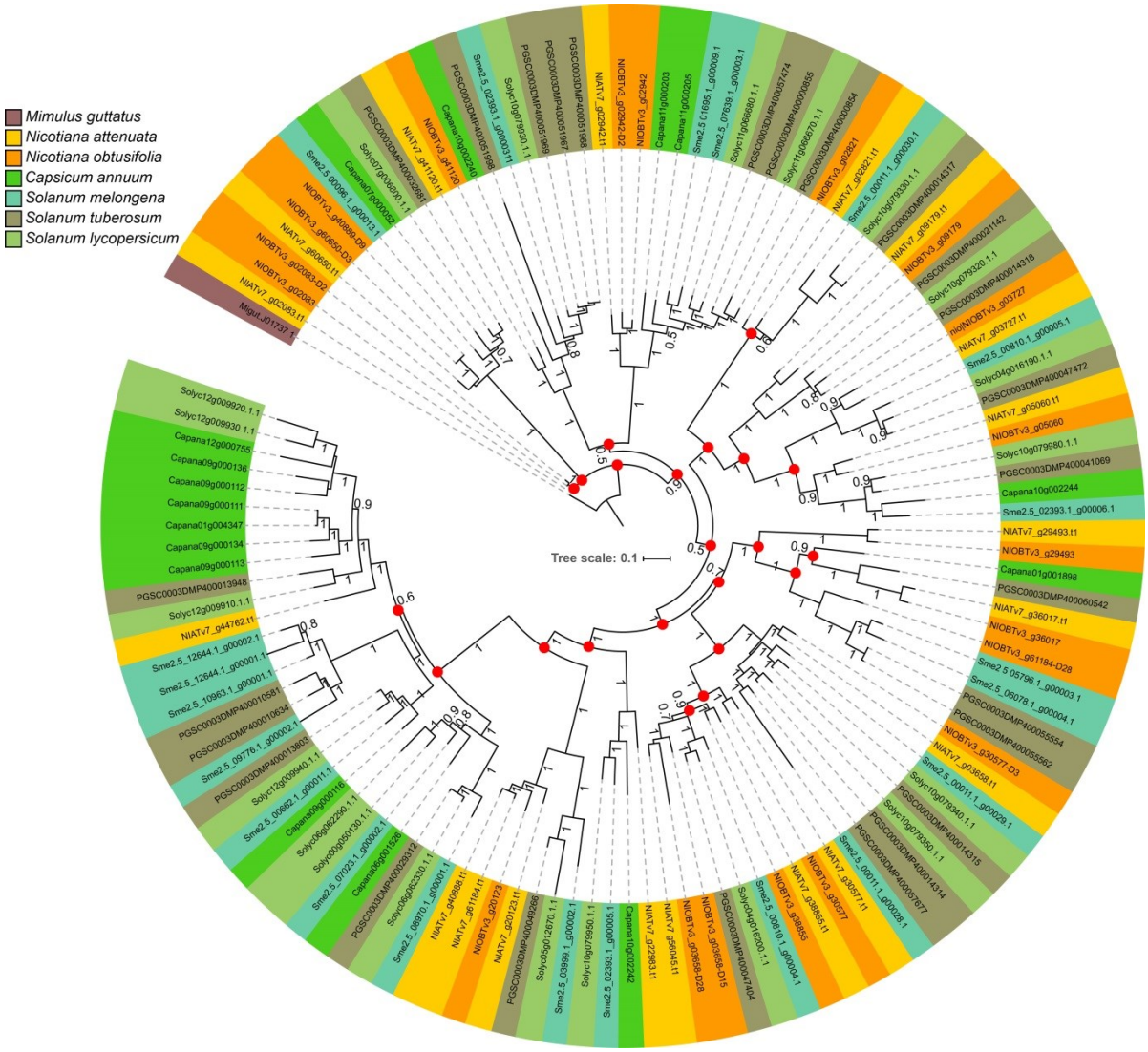
Supplementary Figure S15. Distribution of 4dTV is consistent with the hypothesis that *Nicotiana* and *Solanum* share a genome-wide duplication event. The distributions of fourth-fold degenerate sites (4dTV) between duplicated paralogs within and orthologs between genomes are shown. The comparison *N. attenuata* with tomato reflects the speciation events between *Nicotiana* and *Solanum*. Within genome comparisons shows the divergence between duplicated gene pairs and thus reflects genome-wide duplication events.



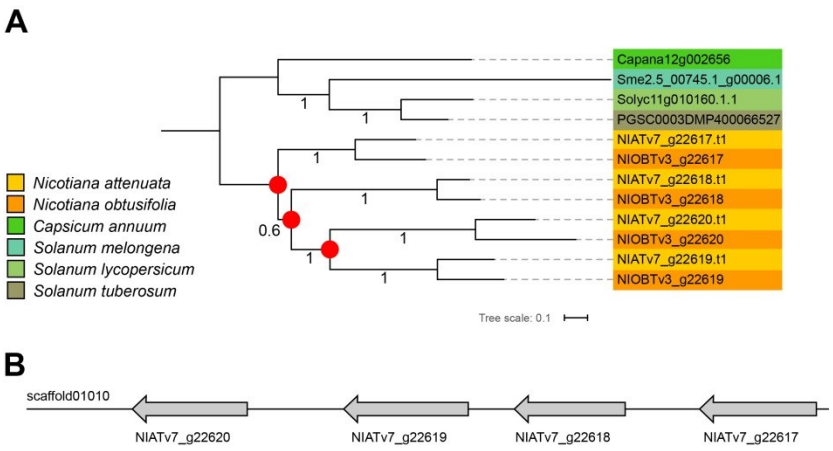
Supplementary Figure S16. Circos plot of the 12 assembled pseudo-chromosomes. Ribbons indicate the 22 syntenic blocks, each of which contains at least 20 genes. **A)** Ka/Ks of genes between *N. attenuata* and *N. obtusifolia*. Red dots indicate the value greater than 1. **B**) *M. sexta* oral secretion induced gene expression changes in *N. attenuata* leaves. Each bar indicates log2 fold change of each gene. **C**) heatmap shows 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.



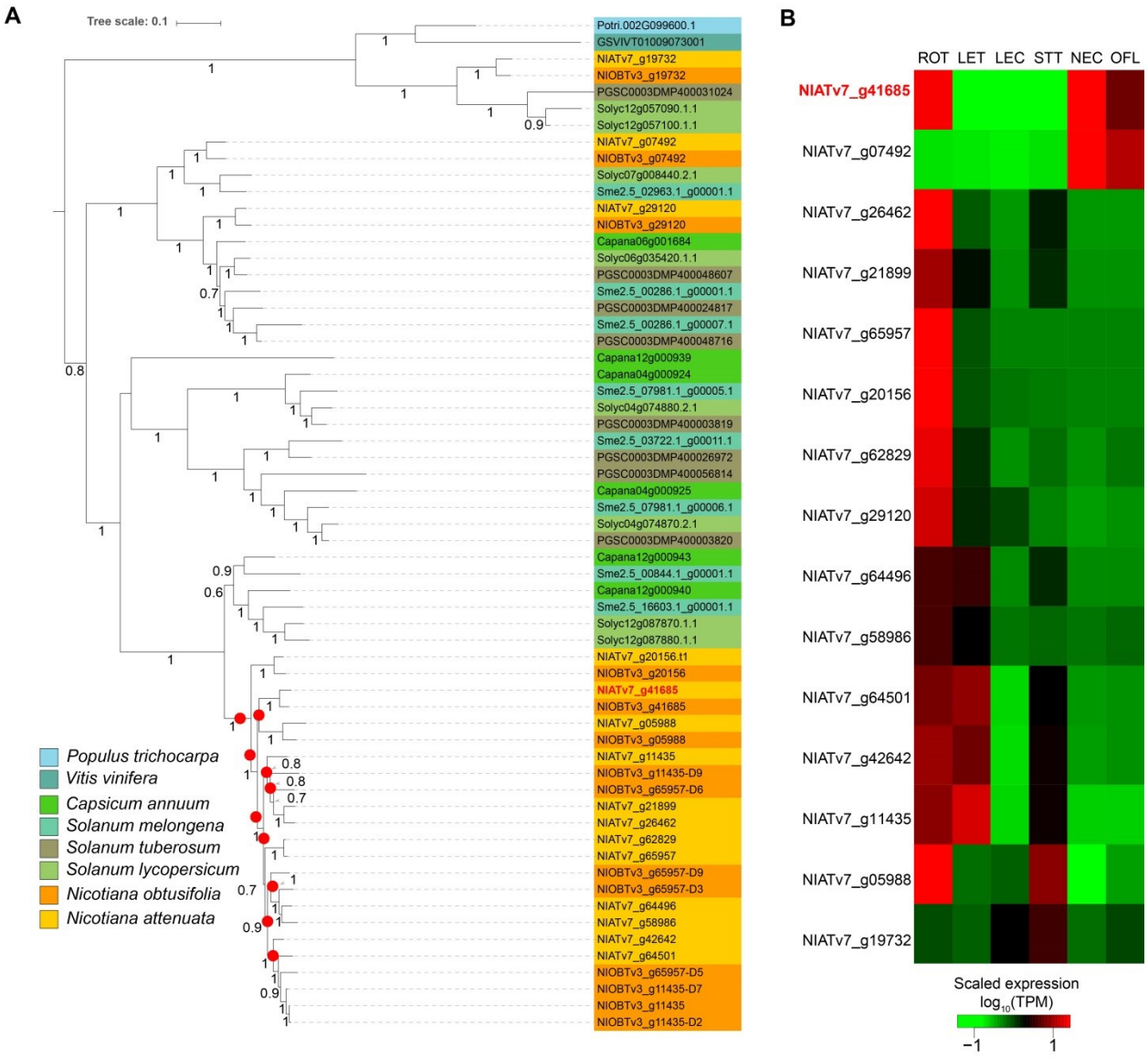
Supplementary Figure S17. Expansion of SLF gene family in Solanaceae. Phylogenetic diagram showing the rapid gene duplication (colors) and loss (grey) of *SLF* subfamily members in the Solanaceae. The sequence from *Pyrus pyrifolia* was used as outgroup. The number at each branch indicates the approximate 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 grey regions. Nodes with only color but no gene ID indicate the gene loss detected in the given species/clade.



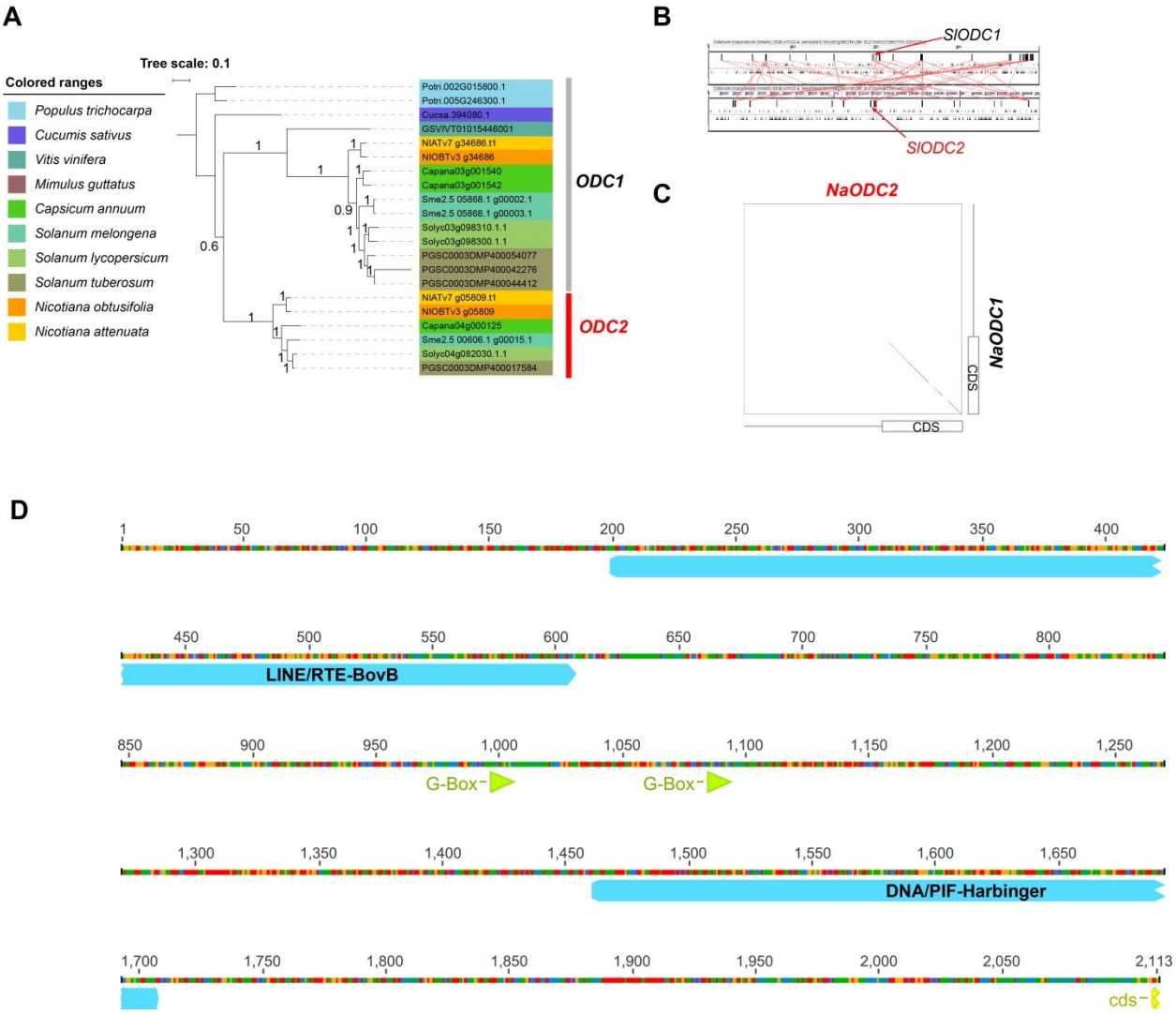
Supplementary Figure S18. Expansion of the *Zeatin O-glucosyltransferase-like* gene family in Solanaceae. Phylogenetic tree showing the expansion of the family. Numbers at the branches refer to the approximate Bayes branch supports. Red circles indicate the detected duplication events shared among Solanaceae species.



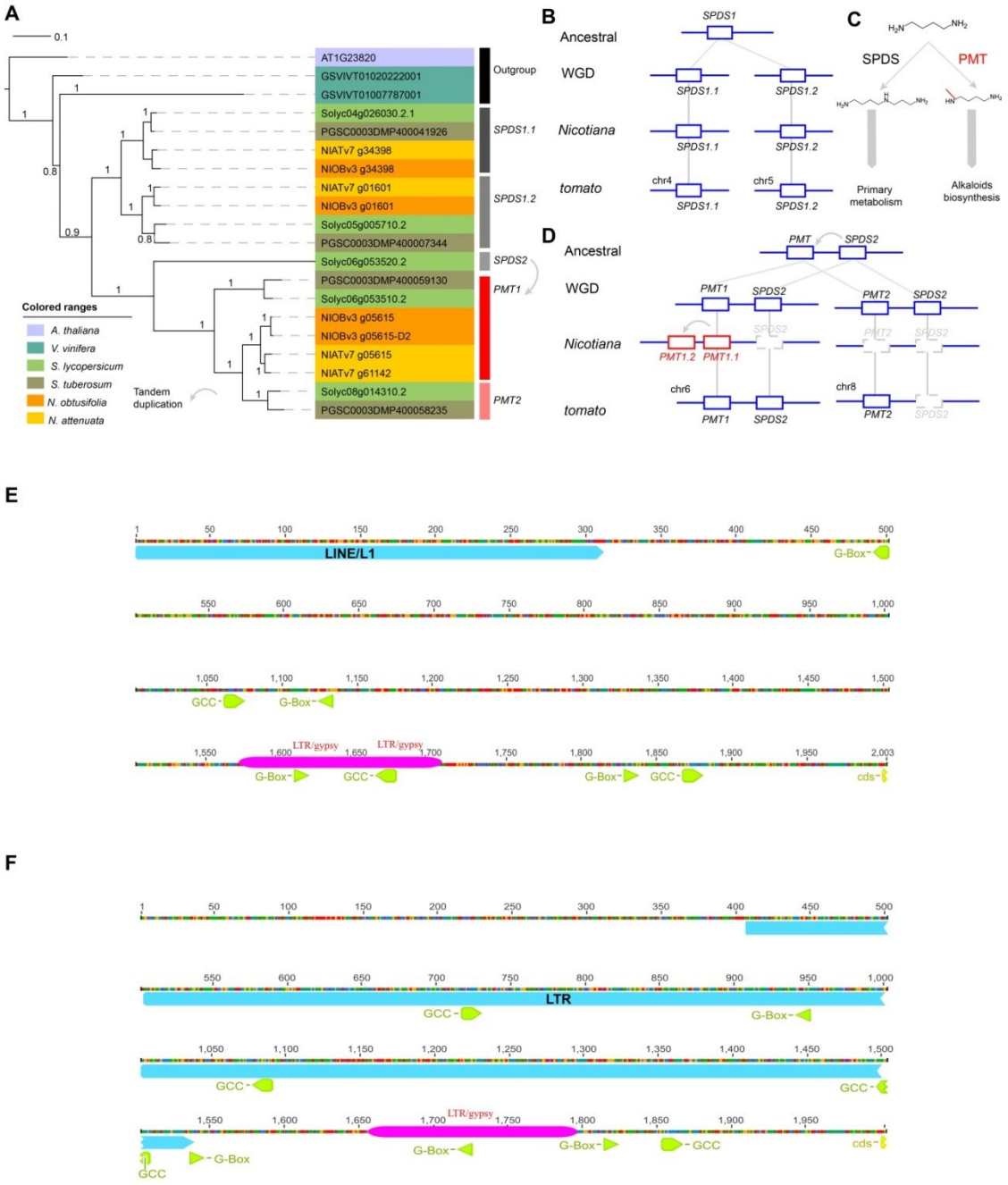
Supplementary Figure S19. Expansion of a disease resistant gene family (*RPM1*-like) in *Nicotiana*. **A)** phylogenetic tree of the gene family. This gene family was found specific to Solanaceae plants. While no duplication was found in other Solanaceae species, three duplication events were identified 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*.



Supplementary Figure S20. 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.

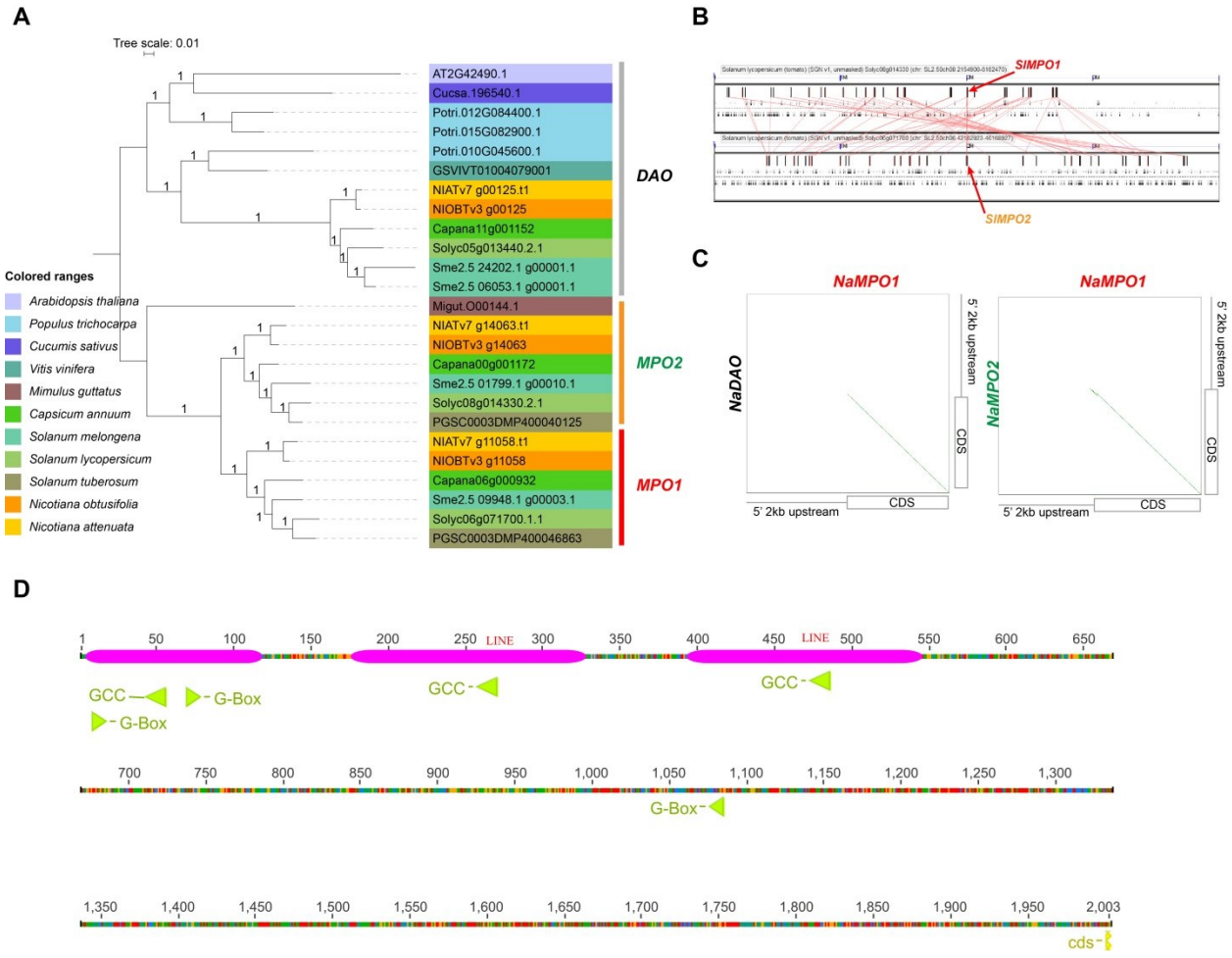


Supplementary Figure S21. Evolution of *ornithine decarboxylases* (*ODC*). A) Phylogenetic tree of *ODC* among different plants. An ancient duplication event occurred before the divergence between *Vitaceae* and *Solanaceae*. The number on the branch shows the approximate Bayes branch support. **B)** syntenic information between ODC1 and ODC2 in tomato. No clear signature of synteny was found. **C)** a dot plot depicts the sequence similarity of CDS and 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 TEs annotated from RepeatMasker.

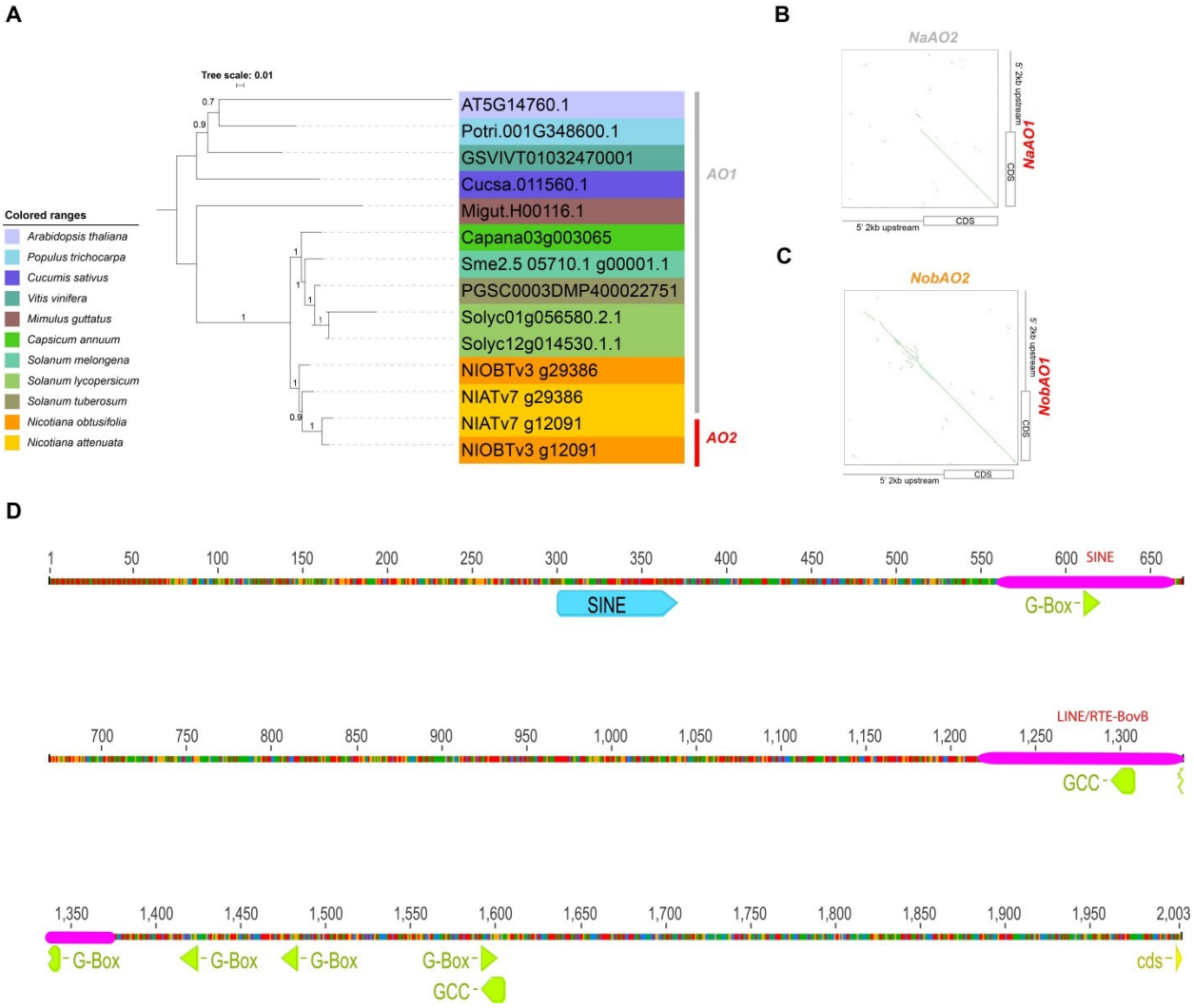


Supplementary Figure S22. Evolution of *N*-putrescine methyltransferases (*PMT*). A) Phylogenetic tree of *PMT* and *SPERMIDINE SYNTHASE* (*SPDS*) in six plant species. The number on the branch shows the approximate Bayes branch support. The closest homolog from *Arabidopsis* and grape were considered

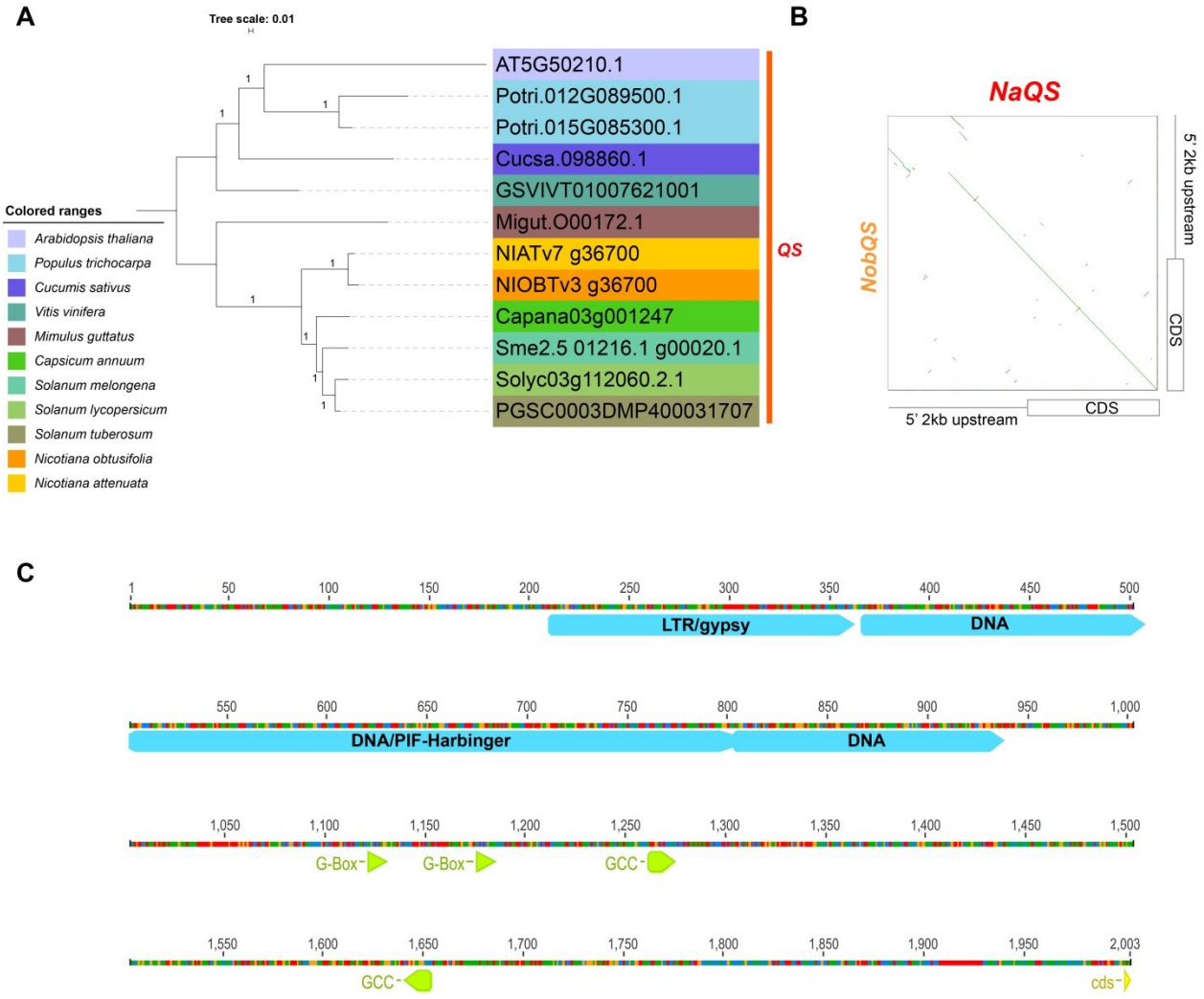
as outgroups. **B)** and **C)** simplified evolutionary model of SPDS and *PMT* in tomato and *Nicotiana*. The syntenic information used to construct these models is from tomato. **D)** a simplified schematic representation of *SPDS* and *PMT* functions. **E) and F)** detailed annotation of TE and transcription 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).



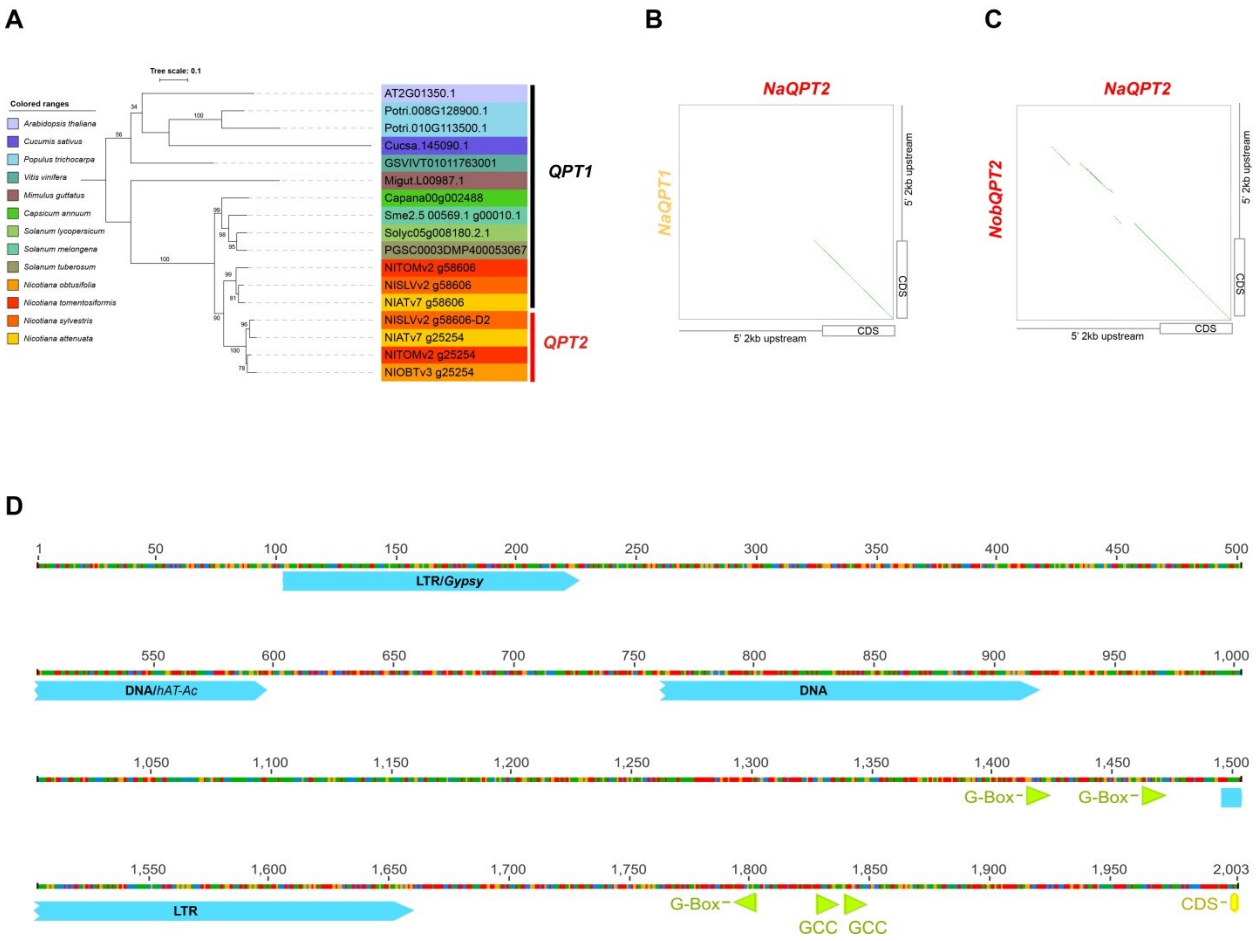
Supplementary Figure S23. Evolution of *N*-methylputrescine oxidases (*MPO*). A) the phylogenetic tree of *MPO* and *DAO* in 11 plant species. *MPO* evolved from the duplication of DAO, likely before the divergence between Phrymaceae (Lamiales) and Solanaceae. The duplication between *MPO1* and *MPO2* are shared among different Solanaceae species. **B)** the *MPO1* and *MPO2* from tomato are in a syntenic block. **C)** two dot plots depict the sequence similarity of protein coding sequences and 2kp upstream regions between *MPO1* and *DAO* (left) and between *MPO1* and *MPO2* 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).



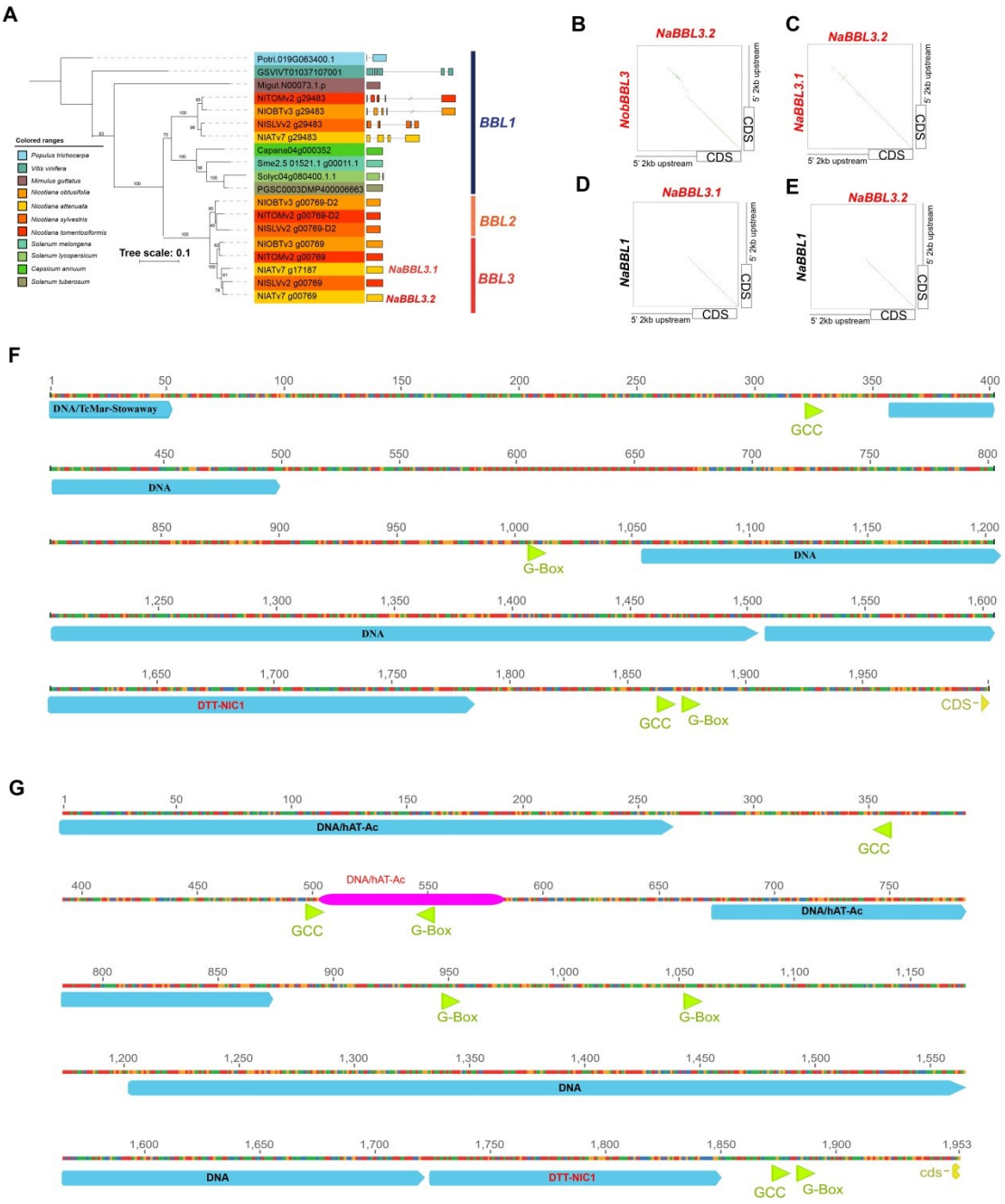
Supplementary Figure S24. Evolution of *aspartate oxidases* (*AO*). A) Phylogenetic tree of *AO* among 11 plants. Both *N. attenuata* and *N. obtusifolia* have two copies of *AO*. The number on the branch shows the approximate Bayes branch support. **B) and C)** the dot plot of CDS sequence together with 2kp 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).



Supplementary Figure S25. 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 2kp 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.

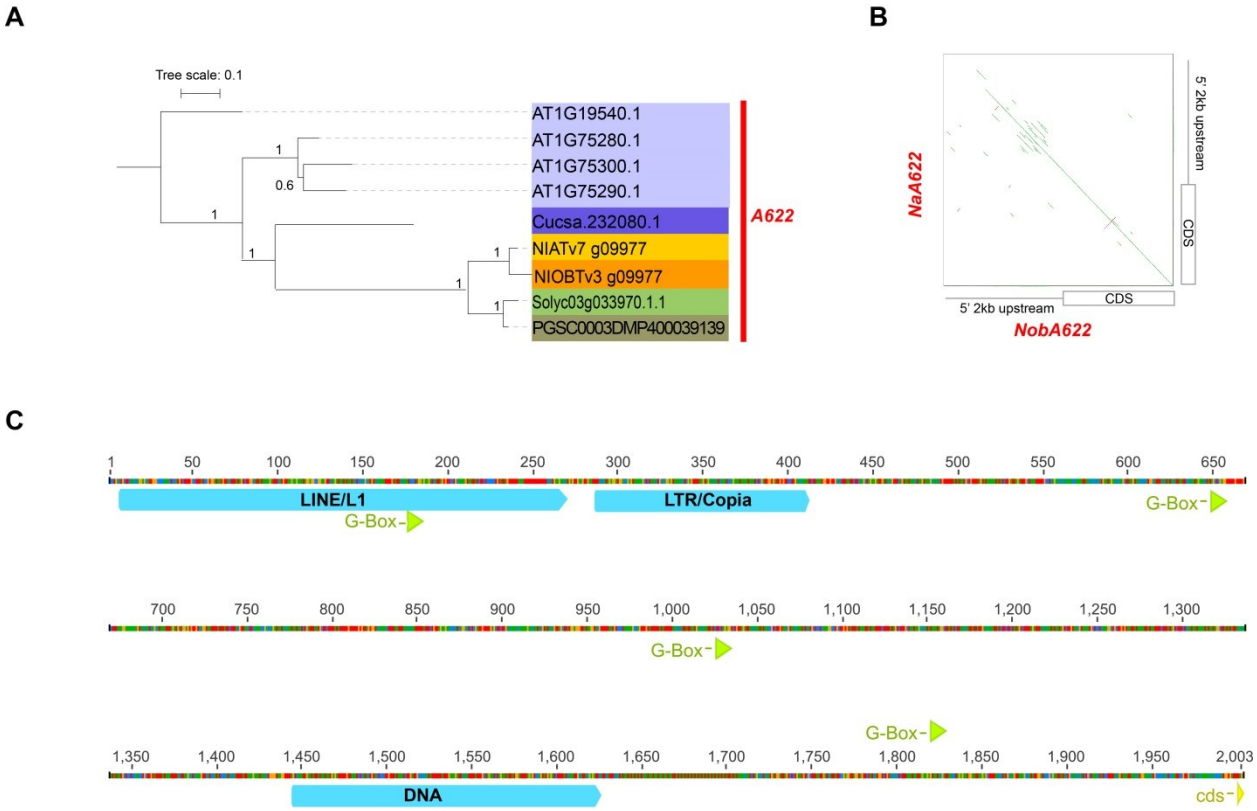


Supplementary Figure S26. Evolution of *quinolinic acid phosphoribosyl transferases* (*QPT*). A) the phylogenetic tree of *QPT* among 13 plants. All *Nicotiana* species have two copies of *QPT,* except *N. obtusifolia*, which is likely due to incomplete genome assembly. The number on the branch shows the bootstrap value**. B and C)** dot plots depicting the sequence similarity of CDS sequence together with 2kp upstream region of *QPT1 and QTP2* 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.



Supplementary Figure S27. Evolution of *berberine bridge enzyme-like proteins* (*BBL*). A) Phylogenetic tree of *BBL* 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 2kp upstream region between *NaBBL3.2* and *NoBBL3* (B), *NaBBL3.1* and *NaBBL3.2* (C), *NaBBL3.1*and *NaBBL1* (D), *NaBBL3.2* and *NaBBL1* (E). **F and G)** detailed annotation of TE and transcription factor

binding motifs of *NaBBL3.1* (F) and *NaBBL3.2* (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 *NaBBL3.1* and *NaBBL3.2* have an insertion of the DTT-NIC1 MITE insertion.



Supplementary Figure S28. Evolution of *A622* protein. A) a subclade tree of the *A622* super family 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 2kp 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).

6. Supplemental Tables.

Supplementary Table S1. Summary of filtered (duplicate removal, quality clipping) WGS sequencing data used for the assemblies of *N. attenuata* and *N. obtusifolia*.

	Nicotiana attenuata (2.5Gb)							*Nicotiana obtusifolia (1.4Gb)*		
Library type	PE180	PE250	PE600	PE950	MP5500	MP20000	SR454	PE480	PE1050	MP3500
Number paired reads	354,327,662	254,673,142	152,657,090	92,178,308	80,509,574	2,005,658	-	478,531,366	222,388,220	108,153,562
Number non paired	103,391,483	30,351,121	68,691,924	77,197,196	-	-	25,188,020	-	-	-
Avg. length [bp]	87.5	97.9	91.7	88.6	75.0	75.0	644.1	91.3	77.6	94.6
Length sum [Gb]	40.03	27.91	20.30	15.01	6.04	0.150	16.22	43.70	17.26	10.23
Approx. seq. coverage	16.7	11.6	8.5	6.3	2.5	0.1	6.8	31.2	12.3	7.3
Approx. frag. coverage	**13.3**	**13.3**	**19.1**	**18.2**	**92.3**	**8.4**	**6.8**	**82.0**	**83.4**	**135.2**

Supplementary 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.

Groups		Complete					Partial				
		Average	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Average**	**Group 1**	**Group 2**	**Group 3**	**Group 4**
N. attenuata	#Prots	208	51	44	52	61	243	63	54	61	65
	%Completeness	**83.87**	77.27	78.57	85.25	93.85	**97.98**	95.45	96.43	100	100
	#Total	463	95	86	121	161	646	145	135	167	199
	Average	2.23	1.86	1.95	2.33	2.64	2.66	2.3	2.5	2.74	3.06
	%Ortho	63.94	52.94	54.55	71.15	73.77	75.31	66.67	74.07	78.69	81.54
N. obtusifolia	#Prots	211	52	45	55	59	241	65	53	61	62
	%Completeness	**85.08**	78.79	80.36	90.16	90.77	**97.18**	98.48	94.64	100	95.38
	#Total	421	96	78	113	134	573	121	117	169	166
	Average	2	1.85	1.73	2.05	2.27	2.38	1.86	2.21	2.77	2.68
	%Ortho	56.87	51.92	42.22	65.45	64.41	70.54	53.85	66.04	88.52	74.19
N. sylvestris	#Prots	195	49	39	50	57	234	60	52	60	62
	%Completeness	**78.63**	74.24	69.64	81.97	87.69	**94.35**	90.91	92.86	98.36	95.38
	#Total	402	85	73	102	142	588	124	118	157	189
	Average	2.06	1.73	1.87	2.04	2.49	2.51	2.07	2.27	2.62	3.05
	%Ortho	60.51	42.86	61.54	64	71.93	73.93	60	71.15	85	79.03
N. tomentosiformis	#Prots	164	51	35	39	39	188	57	42	44	45
	%Completeness	**66.13**	77.27	62.5	63.93	60	**75.81**	86.36	75	72.13	69.23
	#Total	321	84	62	88	87	441	113	88	121	119
	Average	1.96	1.65	1.77	2.26	2.23	2.35	1.98	2.1	2.75	2.64
	%Ortho	56.71	41.18	48.57	66.67	74.36	72.34	61.4	66.67	81.82	82.22
S. tuberosum	#Prots	208	51	43	51	63	239	63	51	60	65
	%Completeness	**83.87**	77.27	76.79	83.61	96.92	**96.37**	95.45	91.07	98.36	100
	#Total	392	84	76	98	134	512	108	102	137	165
	Average	1.88	1.65	1.77	1.92	2.13	2.14	1.71	2	2.28	2.54
	%Ortho	53.85	41.18	53.49	54.9	63.49	65.27	47.62	62.75	70	80
S. lycopersicum	#Prots	215	55	46	54	60	238	63	52	61	62
	%Completeness	**86.69**	83.33	82.14	88.52	92.31	**95.97**	95.45	92.86	100	95.38
	#Total	391	88	73	104	126	497	112	98	136	151
	Average	1.82	1.6	1.59	1.93	2.1	2.09	1.78	1.88	2.23	2.44
	%Ortho	47.44	34.55	43.48	53.7	56.67	58.4	49.21	55.77	60.66	67.74

Supplementary Table S3. Repeat content among six solanaceous genomes.

TE class	TE type	*S. tuberosum* (Mb)	*S. lycopersicum* (Mb)	*N. obtusifolia* (Mb)	*N. tomentosiformis* (Mb)	*N. sylvestris* (Mb)	*N. attenuata* (Mb)
DNA	DNA	10.3	11.9	36.2	66.5	69.4	70.0
	DNA/CMC-EnSpm	7.5	9.4	11.7	7.6	8.6	12.4
	DNA/Dada	0.1	0.0	2.9	5.5	7.8	7.7
	DNA/MULE-MuDR	3.3	3.2	0.4	0.6	0.4	0.6
	DNA/MuLE-MuDR	1.4	5.0	3.2	5.5	9.7	9.5
	DNA/PIF-Harbinger	5.6	4.3	4.7	4.8	5.8	6.0
	DNA/TcMar-Pogo	0.4	0.4	-	-	-	-
	DNA/TcMar-Stowaway	9.2	4.1	2.7	3.9	4.0	3.8
	DNA/hAT	0.1	-	0.1	0.2	0.4	1.1
	DNA/hAT-Ac	7.5	4.1	9.0	11.1	12.5	12.7
	DNA/hAT-Charlie	0.1	-	-	-	-	-
	DNA/hAT-Tag1	1.7	1.2	1.2	2.0	1.5	2.0
	DNA/hAT-Tip100	4.6	2.6	1.5	2.3	2.3	2.2
LINE	LINE	-	-	11.7	14.6	19.2	22.9
	LINE/L1	10.6	7.9	18.2	12.0	18.3	24.9
	LINE/L1-Tx1	-	-	0.0	0.0	0.5	0.6
	LINE/L2	-	-	0.5	0.2	0.5	0.5
	LINE/R1	-	-	0.2	0.3	0.4	0.4
	LINE/RTE-BovB	5.1	3.3	8.6	13.0	15.4	12.4
LTR	LTR	0.1	-	83.1	133.7	129.4	132.5
	LTR/Caulimovirus	3.7	2.4	12.5	1.5	2.4	1.1
	LTR/Copia	29.1	48.8	42.6	61.5	87.1	131.0
	LTR/ERVL	-	-	0.0	0.9	4.1	4.0
	LTR/Gypsy	**244.6**	**318.0**	**519.7**	**864.6**	**1171.4**	**1201.5**
Others	RC/Helitron	2.4	0.6	3.1	0.0	3.0	3.4
	Retrotransposon	3.8	3.5	8.1	3.4	12.6	14.8
	SINE	0.7	0.5	1.3	13.5	2.5	2.5
	Satellite	3.8	1.0	5.9	1.9	3.2	4.1
	Unknown (Mb)	46.4	37.1	3.2	5.1	9.1	9.5
TE size (Mb)		**402.2**	**469.2**	**792.3**	**1236.2**	**1601.6**	**1694.2**
Genome size (Mb)		676.3	737.6	1223.2	1642.4	2047.5	2090.5
Percentage of TE in the genome (%)		**59.5**	**63.6**	**64.8**	**75.3**	**78.2**	**81.0**

Supplementary Table S4. Information on samples used for RNA-seq.

Tissue	Treatment / development stage	Library ID	# raw reads	# clean reads	% uniquely mapped reads	Read length
Roots	*M. sexta* OS induced on leaves	NA1498ROT	327,772,944	317,843,200	93.77	50bp
Leaves	Control	NA1717LEC	61,531,550	16,430,982	93.29	100bp
	M. sexta OS locally induced	NA1500LET	328,071,888	317,905,162	91.94	50bp
Seeds	Smoke solution treated	NA1501SES	51,423,280	49,120,770	90.53	50bp
	Water treated	NA1502SEW	75,944,970	73,525,266	90.31	50bp
	Dry	NA1503SED	63,463,542	61,285,870	89.01	50bp
Stems	*M. sexta* OS treated on leaves	NA1504STT	72,473,514	70,575,710	94.70	50bp
Corollas	Early developmental stage	NA1505COE	44,064,054	23,081,618	95.14	100bp
	Late developmental stage	NA1515COL	41,110,650	17,017,608	94.71	100bp
Stigmas	Mature	NA1506STI	39,281,658	17,878,354	94.88	100bp
Pollen tubes	Germinated on pollen germination medium	NA1507POL	41,692,244	22,451,272	96.83	100bp
Styles	Without pollination	NA1508SNP	45,428,336	21,445,272	94.64	100bp
	Pollinated with non-self pollen	NA1509STO	39,071,214	18,816,496	94.29	100bp
	Pollinated with self-pollen	NA1510STS	50,505,064	23,107,862	95.20	100bp
Nectaries	Mature	NA1511NEC	55,777,620	23,403,620	92.92	100bp
Anthers	Mature	NA1512ANT	41,480,422	18,767,454	96.22	100bp
Ovaries	Mature	NA1513OVA	39,608,326	17,426,638	94.42	100bp
Pedicels	No treatment	NA1514PED	41,520,690	19,508,002	95.14	100bp
Flowers	Fully opened	NA1516OFL	43,791,980	18,550,688	94.87	100bp
	Buds	NA1517FLB	45,864,746	22,167,496	95.47	100bp
Leaves	Leave samples that were collect at different time points	NA1821CTN	37,692,054	36,884,606	95.44	100bp

Supplementary Table 5. Genomes of 11 plant species used for comparative genomic analysis.

Species	Version	# of gene models	URL	Reference
N. attenuata	r2.0	33,449	http://nadh.ice.mpg.de	This study
N. obtusifolia	v1.0	27,911	http://nadh.ice.mpg.de	This study
A. thaliana	TAIR 10	27,416	http://phytozome.jgi.doe.gov/arabidopsis	The Arabidopsis Genome Initiative. 2000 [44]
C. annuum	v2.0	35,336	http://peppersequence.genomics.cn/page/species/download.jsp	Kim et al. 2014 [45]
C. sativus	v1.0	21,503	http://phytozome.jgi.doe.gov/cucumber	Huang et al. 2009 [46]
M. guttatus	v2.0	28,140	http://phytozome.jgi.doe.gov/mimulus	Hellsten et al. 2013 [47]
P. trichocarpa	v3.0	41,335	http://phytozome.jgi.doe.gov/poplar	Tuskan et al. 2006 [48]
S. lycopersicum	ITAG2.3	34,727	http://phytozome.jgi.doe.gov/tomato	Tomato Consortium. 2012 [22]
S. melongena	v2.5.1	42,035	ftp://ftp.kazusa.or.jp/pub/eggplant/	Hirakawa et al. 2014 [49]
S. tuberosum	v3.4	35,119	http://phytozome.jgi.doe.gov/potato	Xu et al. 2011 [50]
V. vinifera	Genoscope 12X	26,346	http://phytozome.jgi.doe.gov/grape	Jaillon et al. 2007 [51]

Supplementary Table S6. Annotation of nicotine biosynthesis genes and number of motifs in 2kb upstream region. TPM: transcript per million.

	Gene ID in *N. attenuata*	*Gene name*	# of G-box motifs	# of GCC motifs	Root expression (TPM)	Reference sequences in database	References
Nicotine biosynthesis copies	NIATv7_g12091	*AO2*	5	2	3640.36	DW001381	[52]
	NIATv7_g36700	*QS*	2	2	3311.29	AF154657	[52]
	NIATv7_g25254	*QPT2*	3	2	2816.12	AJ748263.1	[53]
	NIATv7_g05809	*ODC2*	2	0	3866.97	AF127242.1	[52]
	NIATv7_g61142	*PMT1.1*	4	3	4457.42	D28506.1	[54]
	NIATv7_g05615	*PMT1.2*	4	4	1806.61	D28506.1	[54]
	NIATv7_g14063	*MPO1*	3	3	454.52	AB289456.1	[55]
	NIATv7_g17187	*BBL3.1*	2	2	275.14	AB604219.1	[56]
	NIATv7_g00769	*BBL3.2*	4	3	956.46	AB604219.1	[56]
	NIATv7_g09977	*A622*	4	0	2258.39	AB445396.1	[57]
Ancestral/non-nicotine copies	NIATv7_g00125	*DAO*	1	4	12.01	XM_009602560	-
	NIATv7_g01601	*SPDS1*	1	0	226.22	NM_001302585.1	[58]
	NIATv7_g11058	*MPO2*	2	0	12.86	AB289457	[55]
	NIATv7_g29386	*AO1*	3	0	8.21	XM_009759430	-
	NIATv7_g29483	*BBL1*	0	0	7.43	AB604221	[56]
	NIATv7_g34686	*ODC1*	0	2	11.27	XM_009771959	-
	NIATv7_g58606	*QPT1*	1	1	29.29	AJ748262.1	[53]

Supplementary Table S7. Primers used for validating 2kb upstream sequence region 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 other four genes, likely due to the presences of repetitive sequences.

Gene ID	Forward Primer	Reverse Primer	PCR amplification
NIATv7_g00125	CTCCATCAAAACTGAAGATCCAGTAGAG	GGGGAAGAAACCGTCGCCTTTTCCTGAG	One step amplification
NIATv7_g01601	GTTGGTGTTTATAGAAATATGAGAACCAAC	CGTTGTCGTTGTTGTTGTGGTTGGCTGC	
NIATv7_g04912	CATGAGGGGGAGACATGACAAACAGTTGAG	GCAGCGTCTACACAACAACCTAGGGCCGGC	
NIATv7_g05615	ATCCATCCGCTCGTCCATTTCTCTCATG	TAGAGCCATTTGTGTTGGTAGATATGAC	
NIATv7_g05809	ATCAGACCAGACAAATTAGCTTATGCGGC	GAATGGCCGCCGGGTTCAACCCGGAAACG	
NIATv7_g09978	CTCCGATTTTGGTGACATGAGTTTACCTGC	CAAGGGCTGCTCAACTTCAGACTTCATGC	
NIATv7_g12091	GAGTTGGAGTGGGCTAAAGTTCATGCC	CTGTCCGCATCCTGAAGCGATACCAG	
NIATv7_g14063	CCCCGCTTGGTCCCATTAGGCCTATTGG	GTGCCGTCACTTTCTGTTTAGTAGTGGC	
NIATv7_g20235	CATGAATACAGTAGCAAAAATAGGCGTG	CTCCGAGAAAGAGAAGGTTGCATTAACG	
NIATv7_g20333	CTATCTCGCTCGAAATCGAGAACTTTG	GTCCTACAATGGCTTCTACGGGAGAAAC	
NIATv7_g21863	CTATTTTCCCTGTACTCAGCTTTAGTAACG	CCATTTTCCACTGGCGTCCCCTTCAGATC	
NIATv7_g25254	CGGTTAATCGATTGTCAAATATGATTTCCTAC	CACTGTTGCAGTGAAAGGAATAGCTC	
NIATv7_g28218	GCGGAAAGGATTCCGACAATAGGGGAAATC	GACTCCATGATTTTCCCCATGCTAATGAGC	
NIATv7_g29386	ATGCCTGACAGGTAAGAACATATAGGTAC	CCGCTTCCTGAAGCGATACCAGTTGC	
NIATv7_g29483	GCTAATCAAAGCAACTACACAAGAGTTG	GAATTATGAGCAGAAGCTGAAGAAACATG	
NIATv7_g32622	GATATCCGGTGGTTTAGGTGGAATGGAAG	GCAACAGTTACCAATCGCTTCTTCTCCGC	
NIATv7_g34686	GTACACCGGAGAAAGTATACTCCTAATAC	GGGAACATGACTCTTACCGATAGCTGTAG	
NIATv7_g36700	GACACAATCACGCACAGAGGCCAGAGCAC	GAAGATTTCATGACTAAATTTGCGGCATC	
NIATv7_g42088	CAGATGGATTTTACAATGCCTTTTTGTTG	CAAGGAGATTAAAATCAGAGAAAGAGAAG	
NIATv7_g58606	CTCCCAACTCCAGAATCACCATACG	GTAATTGCATGAGGATGCACTATTGCAGTG	
NIATv7_g61142	GCTGGGGCTGGAGACTAAGGAACCAGAC	GCCATTTGTGTTGGTAGATATGACTTCC	
NIATv7_g14350	GTGAAGTACGGTGTATAGGCATGGTGAC	GATGATGCAAGCGCATGCT	1st nested PCR
	GTGAAGTACGGTGTATAGGCATGGTGAC	GTTGGTGAACTTCGCTCTTAGAGGAAGTC	2nd nested PCR
NIATv7_g00769	CAGTCCACACACATTTCGGCAGGTG	CTTCATTCCACTGTGCTAGATACTGAAT	1st nested PCR
	CAGTCCACACACATTTCGGCAGGTG	CTTGCTGTTTGGTAGGATAATGAGG	2nd nested PCR
NIATv7_g11058	GTCATTATCTTGTCTTTTTCTTCTCATCTC	AAGGATGGCATGTATGAGCTCTTG	1st nested PCR
	CATATTGCTGGTGCATGATATCAC	GGAGGAGTCACCTTGTGCAAAGTTGCGGC	2nd nested PCR
	GGATTGAAAGTAGTATATTTTTGTG	GGAGGAGTCACCTTGTGCAAAGTTGCGGC	3rd nested PCR
NIATv7_g34398	TGGGATATTAAAATTCAATATACCCAGTTTGTAG	GTAAGGAGATTCATTATTGTTATTCGCTTCC	1st nested PCR
	TGGGATATTAAAATTCAATATACCCAGTTTGTAG	CTTGCTTTTGTACTGACCCCTTTGAG	2nd nested PCR left
	GACTAAAACTTCTACAAATCGTCTGC	GTAAGGAGATTCATTATTGTTATTCGCTTCC	2nd nested PCR right
NIATv7_g17187	CTTACACGTACGGTAAGAGGCTGAGT	CCTGCTGCTTGGTAGGATAATGAAC	1st nested PCR
	GAACTTCTTGTTTATACTTATGACAAG	CGGACTAATAAGGAAAGTGTGTCATC	2nd nested PCR left
	GAAACTGTTTCTACGGCTCAAAAAA	CCTGCTGCTTGGTAGGATAATGAAC	2nd nested PCR right
NIATv7_g09977	CTTTCGCTTTAGAGATGGATTCACTCTT	GTAGCCTGTGCCTCCAATTATTAAGATC	1st nested PCR
	CTCAGTTGTTCTAGTGAAGTTGAAG	CACATTTACCTATTATAACATGTAAC	2nd nexted PCR left
	CGATTTTAGCACCACTTTGGCAGCA	GTAGCCTGTGCCTCCAATTATTAAGATC	2nd nested PCR right

Supplementary 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 provide in the last column. The positions indicate the seeds of predicted miRNA targeting sites.

GeneID	Start	End	miRNA	Seed length	GeneName	Root specific expression	Overlap with TE
NIATv7_g00769	-1937	-1929	NIATTr2_miR77253p.2	8	*BBL3.2*	YES	rnd-5_family-2472: DNA/hAT-Ac
NIATv7_g00769	-1228	-1219	NIATTr2_miR1429.5p	9	*BBL3.2*	YES	-
NIATv7_g00769	-1233	-1225	NIATTr2_miR6164	8	*BBL3.2*	YES	-
NIATv7_g05615	-36	-28	NIATTr2_miR7504	8	*PMT1.2*	YES	-
NIATv7_g05615	-1499	-1491	NIATTr2_miR7711.5p.4	8	*PMT1.2*	YES	rnd-5_family-822: LTR/*Gypsy*
NIATv7_g05615	-737	-729	NIATTr2_miR5163.3p	8	*PMT1.2*	YES	rnd-5_family-58: LTR
NIATv7_g05809	-1865	-1857	NIATTr2_miR1429.5p	8	*ODC2*	YES	-
NIATv7_g05809	-1450	-1442	NIATTr2_miR162	8	*ODC2*	YES	rnd-4_family-757: LINE/RTE-BovB
NIATv7_g05809	-1865	-1857	NIATTr2_miR6161	8	*ODC2*	YES	-
NIATv7_g05809	-106	-95	NIATTr2_miR9487	11	*ODC2*	YES	-
NIATv7_g09977	-977	-968	NIATTr2_miR1066	9	*A622*	YES	-
NIATv7_g09977	-688	-680	NIATTr2_miR6161	8	*A622*	YES	-
NIATv7_g12091	-724	-716	NIATTr2_miR8015.3p	8	*AO2*	YES	-
NIATv7_g12091	-1092	-1083	NIATTr2_miR398.3p	9	*AO2*	YES	-
NIATv7_g14063	-1219	-1211	NIATTr2_miR6444	8	*MPO2*	YES	-
NIATv7_g14063	-780	-772	NIATTr2_miR8015.3p	8	*MPO2*	YES	-
NIATv7_g17187	-1700	-1692	NIATTr2_miR5497	8	*BBL3.1*	YES	-
NIATv7_g25254	-793	-785	NIATTr2_miR6161	8	*QPT2*	YES	-
NIATv7_g25254	-1559	-1551	NIATTr2_miR156.5p	8	*QPT2*	YES	-
NIATv7_g25254	-940	-932	NIATTr2_miR5750	8	*QPT2*	YES	-
NIATv7_g36700	-1075	-1067	NIATTr2_miR6164	8	*QS*	YES	rnd-4_family-1369: DNA
NIATv7_g61142	-36	-28	NIATTr2_miR7504	8	*PMT1.1*	YES	-
NIATv7_g61142	-1042	-1032	NIATTr2_miR1429.5p	10	*PMT1.1*	YES	-
NIATv7_g00125	-1151	-1143	NIATTr2_miR7504	8	*DAO*	NO	-
NIATv7_g01601	-1138	-1130	NIATTr2_miR5750	8	*SPDS1*	NO	-
NIATv7_g04912	-1586	-1578	NIATTr2_miR6444	8	*ADC2*	NO	rnd-3_family-18: LTR/*Copia*
NIATv7_g04912	-1190	-1182	NIATTr2_miR8015.3p	8	*ADC2*	NO	rnd-4_family-175: DNA
NIATv7_g05934	-67	-59	NIATTr2_miR5750	8	*ADC1*	NO	-
NIATv7_g11058	-215	-207	NIATTr2_miR8007.5p	8	*MPO1*	NO	-
NIATv7_g11058	-482	-474	NIATTr2_miR172.3p	8	*MPO1*	NO	-
NIATv7_g29386	-1738	-1730	NIATTr2_miR8667	8	*AO1*	NO	rnd-4_family-391: DNA
NIATv7_g29483	-1830	-1821	NIATTr2_miR1429.5p	9	*BBL1*	NO	-
NIATv7_g29483	-825	-817	NIATTr2_miR408.5p	8	*BBL1*	NO	-
NIATv7_g29483	-1830	-1821	NIATTr2_miR6161	9	*BBL1*	NO	-
NIATv7_g34398	-982	-974	NIATTr2_miR7725.3p.2	8	*SPDS2*	NO	rnd-4_family-1284: DNA/TcMar-Stowaway
NIATv7_g34398	-695	-686	NIATTr2_miR6161	9	*SPDS2*	NO	-
NIATv7_g34686	-1228	-1220	NIATTr2_miR5069	8	*ODC1*	NO	
NIATv7_g34686	-447	-439	NIATTr2_miR6164	8	*ODC1*	NO	rnd-4_family-173: DTT-NIC1
NIATv7_g58606	-1401	-1392	NIATTr2_miR426	9	*QPT1*	NO	rnd-2_family-51: LTR/*Gypsy*
NIATv7_g58606	-1962	-1954	NIATTr2_miR9496	8	*QPT1*	NO	rnd-5_family-192: LTR/*Gypsy*

7. Captions for supplementary datasets S1 to S9

Supplementary dataset S1. Summary information on MITE families in *Nicotiana* and their homologous families in *Solanum* genomes. Different superfamilies are represented by different codes. DTA for *Tc1/Mariner*, DTA for *h*AT, 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 S2. Annotation of smRNA in *N. attenuata*.

Supplementary dataset S3. 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 τ index among all tissues except CTN.

Supplementary dataset S4. 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 S5. 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 S6. KEGG ortholog (KO) annotation of *N. attenuata* genes.

Supplementary dataset S7. List of genes annotated as protein kinases and transcription factors.

Supplementary dataset S8. List of recent duplication events of *N. attenuata* genes and their duplicated copies. Genes that are not duplicated or singletons and the gene annotation 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.

Supplementary dataset S9. Gene families that expanded significantly in *Nicotiana*.

8. References

1. VanBuren, R. *et al.* Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508-11 (2015).
2. Cao, H.Z. *et al.* Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* **3**(2014).
3. Shelton, J.M. *et al.* Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics* **16**(2015).
4. Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes* **5**, 337 (2012).
5. McKenna, A. *et al.* The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
6. Wu, Y.H. *et al.* Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genetics* **4**(2008).
7. Tang, H.B. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol* **16**(2015).
8. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).
9. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**(2013).
10. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765-768 (1996).
11. Kim, S.G. *et al.* Tissue specific diurnal rhythms of metabolites and their regulation during herbivore attack in a native tobacco, *Nicotiana attenuata*. *PLoS One* **6**, e26214 (2011).
12. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
13. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100-3108 (2007).
14. Brooks, A.N. *et al.* Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res* **21**, 193-202 (2011).
15. Ling, Z. *et al.* Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*. *Plant J.* **84**, 228-43 (2015).
16. Jin, Y. *et al.* TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593-3599 (2015).
17. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
18. Diboun, I. *et al.* Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC Genomics* **7**(2006).
19. Gotz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**, 3420-3435 (2008).
20. Xie, C. *et al.* KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res* **39**, W316-W322 (2011).
21. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
22. Sato, S. *et al.* The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).
23. Riano-Pachon, D.M. *et al.* PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* **8**(2007).
24. Page, R.D.M. & Charleston, M.A. From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem. *Mol Phylogenetics and Evol* **7**, 231-240 (1997).
25. Zhang, Z. *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259-63 (2006).

26. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92-4 (2010).
27. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
28. Abascal, F. *et al.* TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* **38**, W7-W13 (2010).
29. Capella-Gutierrez, S. *et al.* trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
30. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-91 (2007).
31. Thorne, J.L. & Kishino, H. Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* **51**, 689-702 (2002).
32. Guyot, R. *et al.* Ancestral synteny shared between distantly-related plant species from the asterid (*Coffea canephora* and *Solanum Sp.*) and rosid (*Vitis vinifera*) clades. *BMC Genomics* **13**(2012).
33. Wu, F.N. & Tanksley, S.D. Chromosomal evolution in the plant family Solanaceae. *BMC Genomics* **11**(2010).
34. Williams, J.S. *et al.* Transcriptome analysis reveals the same 17 S-locus F-box genes in two haplotypes of the self-incompatibility locus of *Petunia inflata*. *Plant Cell* **26**, 2873-2888 (2014).
35. Goldberg, E.E. *et al.* Species selection maintains self-incompatibility. *Science* **330**, 493-495 (2010).
36. Havlova, M. *et al.* The role of cytokinins in responses to water deficit in tobacco plants over-expressing trans-zeatin O-glucosyltransferase gene under 35S or SAG12 promoters. *Plant Cell Environ* **31**, 341-353 (2008).
37. Schäfer, M. *et al.* Cytokinin levels and signaling respond to wounding and the perception of herbivore elicitors in *Nicotiana attenuata*. *J Integr Plant Biol* **57**, 198-212 (2015).
38. Hildreth, S.B. *et al.* Tobacco nicotine uptake permease (NUP1) affects alkaloid metabolism. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18179-18184 (2011).
39. Kato, K. *et al.* Tobacco nicotine uptake permease regulates the expression of a key transcription factor gene in the nicotine biosynthesis pathway. *Plant Physiol.* **166**, 2195-U1500 (2014).
40. Kessler, D. *et al.* Unpredictability of nectar nicotine promotes outcrossing by hummingbirds in Nicotiana attenuata. *Plant J.* **71**, 529-538 (2012).
41. Chintapakorn, Y. & Hamill, J.D. Antisense-mediated reduction in ADC activity causes minor alterations in the alkaloid profile of cultured hairy roots and regenerated transgenic plants of Nicotiana tabacum. *Phytochemistry* **68**, 2465-2479 (2007).
42. Dalton, H.L. *et al.* Effects of down-regulating ornithine decarboxylase upon putrescine-associated metabolism and growth in *Nicotiana tabacum L. J. Exp. Bot.* **67**, 3367-81 (2016).
43. Pandey, S.P. *et al.* Herbivory-induced changes in the small-RNA transcriptome and phytohormone signaling in *Nicotiana attenuata*. *Proc Natl Acad Sci U S A* **105**, 4559-64 (2008).
44. Kaul, S. *et al.* Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).
45. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270-+ (2014).
46. Huang, S.W. *et al.* The genome of the cucumber, *Cucumis sativus L. Nat. Genet.* **41**, 1275-U29 (2009).
47. Hellsten, U. *et al.* Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19478-19482 (2013).
48. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-604 (2006).
49. Hirakawa, H. *et al.* Draft genome sequence of eggplant (*Solanum melongena L.*): the representative *Solanum* species indigenous to the old world. *DNA Res.* **21**, 649-660 (2014).

50. Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-U94 (2011).
51. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-U5 (2007).
52. Shoji, T. *et al.* Clustered transcription factor genes regulate nicotine biosynthesis in tobacco. *Plant Cell* **22**, 3390-3409 (2010).
53. Shoji, T. & Hashimoto, T. Recruitment of a duplicated primary metabolism gene into the nicotine biosynthesis regulon in tobacco. *Plant J.* **67**, 949-959 (2011).
54. Shoji, T. *et al.* Jasmonate induction of putrescine N-methyltransferase genes in the root of Nicotiana sylvestris. *Plant Cell Physiol.* **41**, 831-839 (2000).
55. Shoji, T. & Hashimoto, T. Why does anatabine, but not nicotine, accumulate in jasmonate-elicited cultured tobacco BY-2 cells? *Plant Cell Physiol.* **49**, 1209-1216 (2008).
56. Kajikawa, M. *et al.* Vacuole-localized berberine bridge enzyme-like proteins are required for a late step of nicotine biosynthesis in tobacco. *Plant Physiol.* **155**, 2010-2022 (2011).
57. Kajikawa, M. *et al.* A PIP-family protein is required for biosynthesis of tobacco alkaloids. *Plant Mol. Biol.* **69**, 287-298 (2009).
58. Hashimoto, T. *et al.* Molecular cloning of plant spermidine synthases. *Plant Cell Physiol.* **39**, 73-79 (1998).

Manuscript III

Deep learning rapid evolution of alternative splicing in plants

Zhihao Ling, Thomas Brockmöller, Ian T. Baldwin and Shuqing Xu

To be submitted to Genome Biology

Manuscript for Genome biology

Deep learning rapid evolution of alternative splicing in plants

Zhihao Ling, Thomas Brockmöller, Ian T. Baldwin and Shuqing Xu*

E-mail addresses: ZL: zling@ice.mpg.de, TB: tbrockmoeller@ice.mpg.de, ITB: baldwin@ice.mpg.de, SX: sxu@ice.mpg.de

Running title: rapid evolution of alternative splicing in plants.

* Correspondence: Shuqing Xu, Department of Molecular Ecology, Max Planck Institute for Chemical Ecology. Hans-Knöll-Straße 8 D-07745 Jena Germany.

E-mail: sxu@ice.mpg.de

Phone: +49 (0)3641 57 1122

The authors declare no conflict of interest.

Abstract

Background: Alternative pre-mRNA splicing (AS) is prevalent among all plants and is involved in many interactions with environmental stresses. However, the evolutionary landscape and underlying mechanisms of AS in plants remain unknown.

Results: Analyzing the transcriptomes of six plant species revealed that AS diverged rapidly between closely related species, largely due to gain and loss of AS events among orthologues genes. Comparing AS between closely related species showed that the AS that generates PTC containing transcript are more conserved than the AS that generates non-PTC containing transcripts, although only present a minor proportion of total transcripts generated from AS. Furthermore, among six ultra-conserved AS events across different species, five AS generate transcripts containing premature termination codon (PTC), suggesting AS coupled with nonsense-mediated decay (NMD) plays an important role at post-transcriptional level in plants. To understand the mechanisms underlying the rapid evolution of AS, we analyzed the key AS determinants using a machine learning approach. Among all investigated six plant species, the distance between the authentic splice site and the nearest alternative splice site and the size of exon-exon junctions are the two major determinants for both AltD and AltA, suggesting a conserved mechanism for alternative splicing in plants. Variations of these two determinants significantly contributed to the AS turnover between closely related species in both Solanaceae and Brassicaceae.

Conclusions: Our results revealed rapid gain and loss of AS events in plants, the process of which was likely due to changes of a few key AS determinants associated with splicing sites. The results provide new mechanistic insights on evolution of AS in plants and shed further light on the role of post-transcriptional regulations in plant-environment interactions.

Keywords: alternative splicing, evolution, transcriptome, splicing code, deep learning, nonsense-mediated decay

Background

Due to their sessile lifestyle, plants evolved various mechanisms in response to environmental stresses. Alternative splicing (AS), a mechanism by which different mature RNAs are formed by removing different introns or using different splice sites (SS) from the same pre-mRNA, is known to be important for stress-induced responses in plants [1, 2]. Both biotic and abiotic stresses such as herbivores [3], pathogens [4], cold [5] and salt [6] can all induce global AS changes in plants. The environmental induced AS changes in turn can affect phenotypic traits of plants and may contribute to the adaptation to different stresses [1, 2]. For example, low temperature-induced AS changes of flowering regulator genes affect the flowering time and floral development in *A. thaliana* [7, 8]. The strong association between AS and environmental stimulus predicts that AS might have been involved in adaptation process in plants, thus evolved rapidly.

The function of AS can be divided into two major categories: (i) to expand proteome diversity when different transcript isoforms are translated into different proteins (with different subcellular localization, stability, enzyme activity and etc.) [9-11]; (ii) to regulate gene expression by generating pre-mature termination condone (PTC) containing transcripts that are recognized by nonsense-mediated decay (NMD) and results in transcripts degradation [12-14]. Although it was initially thought to be transcriptional noise, several examples of AS events coupled with PTC were found highly conserved in animals [15, 16] and plants [17-19], indicating that AS coupled with NMD might play an important role at post-transcriptional level.

However, it is unclear, to which extent, AS-coupled NMD are more conserved than AS that generate non-PTC containing transcript.

Comparing to vertebrate, the evolution of AS in plants remains largely unclear. Studies that compared organ transcriptomes from different vertebrate species spanning ~350 million years of evolution showed that the AS complexity differs dramatically between vertebrate lineages, and AS evolved much faster than gene expression [20, 21]. For example, within 6 million years, the splicing profiles of an organ are more similar to other organs of same species than the same organ in other species [20, 21]. In plants, largely due to lack of comprehensive transcriptomic data, such comparative analysis remains lacking. However, several indications suggest the pattern of AS evolution in plants and vertebrate might be similar. For examples, only 16.4% of AS between maize and rice, and 5.4% between *Brassica* and *Arabidopsis* are conserved, respectively [22, 23]. A more recent study further showed that only 2.8% of genes showed conserve AS between two species of mungbeans, *Vigna radiate* and *Vigna angularis* [24]. Furthermore, even between different ecotypes of the same plants, there are dramatic changes of AS [25]. However, such low conservations of AS found between species could also be due to several other confounding effects. For example, it is also known that the levels of gene expression, which are highly associated with AS, also diverge rapidly in plants [26]. It remains unclear whether the observed low conservations of AS were due to rapid expression changes between species. Furthermore, detections of AS are highly dependent on sequencing depth and tissue types that used for generating the transcriptomes [3, 27, 28]. Therefore, systematically controlling different confounding effects are necessary to draw conclusive pictures on the evolutionary pattern of AS in plants.

Assembly of the spliceosome to remove introns and ligate exons is directed by sequence features of the pre-mRNA, however, in comparison to animals, the recognition processes of exonic and intronic regions in plants are poorly understood. In metazoans, it is known that four crucial signals are required for accurate splicing: (i) 5' splice sites (SS), which contains a GU dinucleotide at the intron start surrounded by a longer consensus sequences with less conservation, (ii) 3' SS, which includes an AG at the extreme 3' end surrounded by similar sequences of 5' SS, (iii) a polypyrimidine tract (PPT) and (iv) a branch site (BS) sequence that located ~17-40 nt upstream of the 3' SS [29, 30]. In plants, similar sequence features with small difference at specific positions were found in plants, except BS [10]. Besides, a UA-rich tract in introns was also found to be important for efficient splicing in plants [31-33]. In animals, the regulation of splicing also depends largely on *cis* signals and *trans*-acting splicing factors (SFs) that can recognize the signals [20, 21]. Among different SFs, serine/arginine-rich (SR) proteins are an important SFs family that has been showed to be involved in AS regulation [10, 34-37]. In addition, many splicing regulatory elements (SREs) and RNA binding proteins (RBPs) have been identified in animals using different strategies, and the interactions between these SREs in the pre-mRNA and RBPs were found either promote or suppress the use of a splice sites [38-40]. Plants have much higher number of SR proteins, the number of which is almost double of the number in non-photosynthetic organisms, although the number varies among different plant species [18, 41, 42]. Using computational approaches, more than 200 RBPs and 80 SREs in plants have been identified [43]. However the function of which only a few have been confirmed by experiments [44-47].

In mammals, it was suggested that the emergence of AS is originated from constitutive splicing through the fixation of SREs and the creation of alternative competing SS [48, 49]. It

was also observed that features between alternative spliced exons/introns and constitutive spliced exons/introns are different and these features can be used to accurately predict specific AS type [48, 50]. Furthermore, other factors including secondary and tertiary RNA structures, chromatin remodeling, insertion of transposable elements (TEs) and gene duplication (GD) are also known that may play roles in regulating AS [51-58]. However, to which extend, changes of these factors contributed to the evolutionary landscape of AS in vertebrates remains largely unclear. Recently, a study using millions of synthetic mini-genes with degenerated subsequences demonstrated that the likelihood of AS decreases exponentially with the increased distance between the constitutive and newly introduced alternative splicing sites (SS) [59], suggesting sequence changes between constitutive and alternative SS might contribute to the rapid changes of AS between species. In plants, however, the detailed mechanisms that affect the AS remain largely unclear [60]. Although it has been proposed that chromatin landscape changes such as DNA methylation, histone marks, RNA structural features, and SREs are important for regulating AS in plants, experimental evidences remain lacking [60]. A recent study shows that DNA methylation could affect AS in rice [61], indicating changes of DNA methylation might contribute to the variations of AS between species. However, this hypothesis has not been thoroughly tested.

Because AS regulation is a complex process and many factors are involved, computational models are useful tool for identifying key factors and predicting the outcome of splicing. While the Bayesian neural network (BNN) method was developed for decoding the splicing code in mammals [20], deep learning (DL) approach, which refers to methods that map data through multiple levels of abstraction, have been recently shown to surpasses the BNN method [62, 63]. Recently, deep learning methods have been proved to show very promising

power to deal with large, heterogeneous and high-dimensional datasets including predicting DNA and RNA-binding proteins [64] and AS [62, 63].

In this study, we performed comparative analyses on the transcriptomes of both closely and distantly related plant species to study the evolution of AS in plants. To further understand the mechanisms underlying the AS evolution in plants, we applied deep learning approach investigated the determinants of AS and their effects on the AS evolution. Specifically, we aimed to address following questions: 1) what are the evolutionary pattern of AS in plants? 2) Are AS events coupled with NMD more conserved than the regular AS? 3) Which factors are important for the determination of AS in plants; 4) what factors contributed to the rapid turnover of AS in plants?

Results

Genome-wide AS patterns are species-specific in plants

To provide an overview of AS evolution among different plant families, we studied the genome-wide AS pattern of *Arabidopsis thaliana*, soybean (*Glycine max*), tomato (*Solanum lycopersicum*) and wild tobacco (*Nicotiana attenuata*), which have comparable transcriptomic data available from same tissues (roots, leaves and flowers) and represent a wide range of eudicots. The overall distributions of different AS types within each species are consistent with previous studies. In all species, intron retention (IR) and alternative 3' acceptor site (AltA) occupies the two largest portions (Figure S1) [3, 65-67].

To investigate the evolution pattern of AS, we compared AS profiles across selected tissues and species. Because sequencing depth is known to affect the detection of AS, to avoid the heterogeneity of sequencing depths, we randomly subsampled 16 million (the lowest depth

among all samples) unique mapped reads from each sample. All down-stream analyses were based on these equal sequencing depth dataset. Clustering analyses using percent spliced index (PSI), which measure the qualitative differences of AS among samples, showed that different tissues of the same species are more similar to each other than the same tissue from different species (Figure 1A), suggesting AS evolves rapidly in plants. Using qualitative measures of AS by considering the presence and absent of AS (binary) for all one-to-one orthologues genes, same species-specific clustering pattern was found (Figure 1B). Furthermore, consistent patterns were also found when each type of AS was analyzed separately (Figure S2).

To further investigate the evolutionary pattern of AS among closely related species, we analyzed a recently published transcriptome dataset from three Brassicaceae species (*A. thaliana*, *Arabidopsis lyrata* and *Capsella rubella*), each of which have comparable transcriptome data of two tissues (root and shoot) and two treatments (control and cold treated) [68]. Using both quantitative (PSI) and qualitative measures (binary) of AS, a similar species-specific clustering pattern was observed (Figure 1C and D). Interestingly, within same species and same tissue, samples treated with cold stress are clustered together at levels of PSI, consistent with previous studies that stresses can induce genome-wide AS responses [3, 6, 69].

Genome-wide AS regulations diverge faster than gene expression among closely related species.

Species-specific clustering patterns were also reported at gene expression (GE) level among *A. thaliana*, rice and maize [26]. To examine whether species-specific AS clustering is resulted from gene expression divergences, we compared the divergence patterns of AS and GE among transcriptomes of different species. Comparison among species from different plant

families showed that both GE and AS are species-specific clustered (Figure 1A and B, S3A and B). However, when comparing species from the same plant family, such as tomato and *N. attenuata* (Solanaceae), while AS remained species-specific pattern (Figure 1A and B), GE showed tissues-clustering pattern (Figure S3A and B). This suggests that the expression profiles among same tissue from different species are more similar to each other than different tissues from the same species. This pattern is also supported by the expression profiles of tissue samples from the three Brassicaceae species, in which the expression profiles of shoots and roots from different species were clearly separated (Figure S3C and D). These results indicate that AS evolved faster than GE in plants, the pattern of which is similar to animals [20, 21].

Rapid gain and loss of AS among different plants

Species-specific clustering of AS pattern suggests low conservation of AS among species. Overall, among 3857 one-to-one orthologues in the four eudicots that have AS in at least one species, only ~7% of them have AS in all four species, while ~41% of them have species-specific AS. We further investigated the conservation of AS in these genes at exon-exon junction (EEJ) level by performing comparisons between all pair-wise combinations of species. For example, between *N. attenuata* and tomato, 64% (708 out of 1109) of EJJs with AS in *N. attenuata* were found not alternatively spliced in tomato, and 77% (1359 out of 1760) of EEJs with AS in tomato were not alternatively spliced in *N. attenuata*. The rapid change of AS could be results of rapid loss or gain of EEJ between species. To exam the conservation of EEJs, we compared the EEJ structures among pair-wise orthologues genes. In total, 60% of EEJs are conserved in at least two species (while only ~12% for AS), and analysis based on AS events from the most conserved EEJs (found in all four species) showed that 92% of them are specie-

specific (Figure S4A). Similar pattern was found based on the analysis of EEJs that are conserved in at least two species, in which only 10 AS events (0.25%, out of 4,015) were found conserved among all four species (Figure S4B). We also performed the same analysis within the three Brassicaceae species and found 69% of total EEJs are conserved in at least two species (while only 27% for AS). 72% of AS events take place in the most conserved EEJs (found in all three species) are specie-specific (Figure S5A). Furthermore, only ~8% of AS events (1,476 out of 19,170) are conserved among all three species (Figure S5B). The consistent results between divergent species and closely related species suggest that rapid changes of AS took place in each plant species.

To investigate the transition spectrum of AS between each two species pairs, we calculated the transitions among different types of AS. Among the four eudicots, while the transitions among different AS types are rare, gain/loss of AS was found to be the most abundant transition type from all comparisons (Figure 2A-F). Among different AS types, AltA and ES are the most and least conserved AS, respectively. When comparing AS transitions among three closely related species in Brassicaceae, similar patterns are observed (Figure S6A-C). These results suggest that the species-specific AS pattern is not largely due to the rapid changes of EEJs among species, but rapid species-specific gain and loss of AS.

The group of AS that generate PTC-containing alternative transcripts are more conserved than others

Previous studies suggest that many pre-mRNAs undergo unproductive AS, which generate transcripts with in-frame PTCs that coupled to nonsense mediated decay (NMD) in plant [12, 70]. To investigate whether unproductive AS can affect the AS conservation and

contribute to the rapid loss/gain of AS among different plant species, we separate the AS into two groups: (1) PTC+ AS and (2) PTC- AS (details see Materials and Methods). Overall, the portion of PTC+ AS range from 9% - 15% among the four eudicots (Figure S7), suggesting that only a small portion of AS generate PTC containing transcripts. Comparing the levels of conservation between tomato and *N. attenuata*, we found the PTC+ AS is significantly more conserved than PTC- AS ($P < 0.02$, Figure 3A). Furthermore, among nine PTC+ AS of *N. attenuata* which are both conserved and have PTC information in tomato, eight of them (89%) also generate PTC+ transcripts in tomato.

To provide further understanding on PTC+ AS conservation and functional regulations, we extended our analysis by adding the transcriptome data of a very ancient plant species, the spreading earth moss (*Physcomitrella patens*). We focused on the 10 most highly conserved AS events found in all four eudicot plants (Figure S4B) and check whether they can also been found in moss. In total, we found six AS events that also present in moss, indicating these AS events might have evolved since land plants and played essential functions in plants. Interestingly, two of these ultra-conserved AS events were from serine/arginine-rich (SR) genes (*RS2Z33*-like and *RS40*-like), which are part of RNA splicing machinery and the identified AS in *RS2Z33*-like gene was also found in rice [18]. Analyzing the protein coding potential of the transcripts generated by these six conserved AS events showed that five of them resulted in PTC+ transcripts. For example, the AS of *RS2Z33*-like and *RS40*-like genes result in PTC+ alternative transcripts in all five species and likely to be the targets of NMD (Figure 3B). To further investigate whether these PTC+ transcripts are regulated by NMD, we analyzed the available transcriptome data from *A. thaliana* wide type (WT) and NMD-deficient (*lba1* and *upf3-1* double mutant) plants [70]. Among all five PTC+ transcripts in *A. thaliana*, three showed significantly

higher expression in NMD-deficient plants (P < 7e-06, as shown in [70]) including *RS2Z33*-like and *RS40*-like genes (Figure 3B). These results suggest that the mechanism that AS coupled with NMD are more conserved than regular AS and some of them might have evolved long time ago.

Mechanisms involved in determining AS are overall conserved among different plant species

To further understand mechanisms that contributed to the rapid turnover of AS among species, it is necessary to identify the determinate features of AS in plants, which was largely unknown [60]. Because splicing is often mediated by SS, we were interested whether the SS were different between constitutively and alternatively spliced junctions. Comparisons on the SS and their 12 bp surrounding sequences between constitutively and alternatively spliced junctions showed that their SS are overall very similar (Figure S8). Furthermore, we separately identified sequence motifs (12-mer) that enriched in 5' and 3' splice sites (SS) compared to random sequence and found these identified motifs are also highly conserved among studied species (Figure S9). These results indicating changes in SS did not play a key role for the rapid turnover of AS among species.

In addition to SS, it has been showed that distance between regular and the nearest alternative SS, splicing junction size and SS strength are also important for the regulation of different AS types [50, 59, 71]. For different AS types, we compared these features from both constitutively and alternatively spliced junctions. Because exon skipping (ES) events are rare in all species, we only studied the three main AS types (AltD, AltA and IR). The results showed that for a given junction, while the likelihood of both AltD and AltA decreases with distance between regular and alternative SS as well as the distance between regular SS and the nearest

internal GT/AG, the likelihood increases with junction size (Figure 4A and B). Interestingly, although the likelihood of IR in smaller junctions appears bigger than large junctions, no significantly correlation with junction size was found (Figure S10A). Both 5' and 3' SS of junction with IR are significantly weaker than constitutive junction (Figure S10B).

Furthermore, the presence/absence of UA-rich tract, polypyrimidine (PPT) tract, branch site (BS) are also known to be associated with 3' splicing recognition in eukaryotes [30, 32]. We compared the frequency of AltA and IR between junctions of AS gene with and without the presence of UA, PPT tract and BS within 100 bp upstream of 3' SS. We found the frequencies of both AltA and IR are significantly higher in junctions without UA and PPT than junctions with them, while the presence of BS has no significant effect (Figure S11).

It has been hypothesized that *cis*-regulatory elements, including enhancers and silencers near SS are also important for the regulations of splicing. To identify these candidate regulatory elements, we performed *de novel* hexmer motif enrichment analysis by comparing 50 bp sequences from 5' and 3' sides of both donor and acceptor sites between alternatively spliced and constitutively spliced junctions. The results showed that most of the putative enhancer motifs for AS junctions are highly similar to the identified SS. In addition, we also identified several putative silencer motifs (range from 5-10 for AltD and 10-18 for AltA in the five species), which are significantly more enriched in constitutively spliced junctions than AS junctions (Figure S12 A and B).

To evaluate whether these identified features represent the AS determinants, we used a machine learning approach and modeled the different types of AS in each of the studied species. The rational for this approach is that if the features we analyzed represent the AS determinants, using these information, we will be able to predict whether a splicing junction is constitutively or

alternatively spliced. For this, we further extracted information on whether alternative SS would introduce a frame-shift, which would likely result in premature terminate code (PTC), the number of reads that support the junction, which represent levels of expression that is known to be associated with AS, as well as presence and absence of the identified *cis*-motifs. Overall, our models achieved high precision and specificity for both AltD and AltA in all five species (Figure 4C and D, S13A and B), suggesting the features we analyzed can provide sufficient information to discriminate AS junctions from constitutively spliced junctions. However, for IR, the extracted features did not result in a good prediction model (the average AUC is 0.54), indicating that additional undetected factors may have contributed to the determination of IR.

The modeling approach further allows us to compare the relative contributions of each feature to the prediction model. The results showed that for AltD, the nearest alternative 5'SS, junction size and the nearest inter-GT are the top three important features for the prediction among all. In addition, the frame shifts introduced by the nearest alternative 5' SS and nearest GT are also important factors (Table S1). For AltA, the distance to the nearest inter-AG is the top important features for the prediction among all five species. Interestingly, all of the identified putative silencers only had marginal role for the predictions of both AltD and AltA (Table S1). Together, these results showed that the mechanisms regulating AS are overall conserved among studied species.

Changes of AS determinants contributed to the rapid turnover of AS in plants

The highly conserved AS regulation mechanisms among studied species provides a foundation for investigating the mechanisms that contributed to rapid turnover of AS among closely related plant species. Because we didn't find determinants for IR, we only focus on the

evolution of AltA and AltD. We first calculated the divergence of the nearest alternative 5'/3' SS and inter-GT/AG between closely related species (*N. attenuata* versus tomato; *A. thaliana* versus *A. lyrata* and *A. lyarata* versus *C. rubella*). For both AltD and AltA, the levels of conservation decreases with the changes of the distance between constitutive SS and alternative SS in all three pairs of comparisons (Figure 5A and B). In addition, the changes of reading frame introduced by the alternative SS also significantly decreased the conservations of both AltA and AltD (Figure 5C and D). Same pattern was also found for the distance between constitutive SS and nearest inter-GT/AG (Figure 5E-H).

Then we investigated effects of *cis* regulatory elements (CREs) changes: UA-track, PPT and BS (Figure S11 and Table S1). Consistent with the functional roles of these CREs in regulating AltA, while changes of UA, PPT and BS did not affect the conservation of AltD between species (Figure S14A), they significantly reduced the conservation for AltA (Figure S14B).

To further systematically analyze different factors that might affect the conservation of AS, we further constructed the AS evolution model using deep learning method. In addition to the important AS determinants we identified in this study, in the model, we also included several other features that were previously hypothesized to be important for AS conservation between species, such as differences at copy numbers (role of gene duplications), differences at transposable element (TE) insertion within the junction, GC-content and sequence similarity of SS. For AltD, all three between species models achieved significant better prediction than by chance (highest *P*-value = 3e-44), with average precision of 0.63 and specificity of 0.82. In all three models, the top three informative features for the predictions are: distance changes to nearest alternative 5' SS and the nearest inter-GT (Figure S15A and Table S2). For AltA, all

three models also achieved even higher precision and specificity (average 0.70 and 0.85, respectively), and the prediction is significantly higher than by chance (highest *P*-value = 3e-145). In all three models, changes of the nearest inter-AG and alternative 3' SS and the divergence of CREs represent the top three features that contributed to the model predictions (Figure S15B and Table S2).

Interestingly, we found TE insertion is also an important factor that reduced the conservation of both AltD and AltA between *N. attenuata* and tomato but not between any of two Brassicaceae species (Figure S16A and B). This likely due to the difference of TE abundance between *N. attenuata* (~63%) and tomato (~81%) is much higher than the difference between *A. thaliana* (~23%) and *A. lyrata* (28%) [72, 73]. Furthermore, we also included the analyzed the impact of DNA methylation changes between *A. thaliana* and *A. lyrata* using data from [68]. No significant effect was found (Figure S15).

Discussion

In this study, we showed that the divergence of AS profiles is faster than gene expression and rapid gain and loss of AS resulted in lineages specific AS profile in plants. Although AS events that introduced premature termination code (PTC), represent only a small proportion of total AS, they are more conserved than AS without introducing PTC (Figure 3). Several AS coupled with PTC events were found ultra-conserved among highly divergent plants. Using a computational modelling approach, we identified several key determinants of both alternative donor (AltD) and alternative acceptor (AltA) splicing in five different plant species and found changes of these key determinants significantly contributed to the rapid gain and loss of AS between closely related species.

In this analysis, we have observed a dominant species-specific pattern of AS in both diverged and closely related species, suggesting that AS in plants diverges rapidly (Figure 1). Such rapid evolution of AS in plants is similar to the pattern found among vertebrate species that spanning ~350 million years of evolution [20, 21], in which AS is largely segregated by species, while GE is segregated by tissue types [20, 21]. This indicate that the rapid evolution of AS might be universal in among eukaryotes. Interesting, in plants, the evolution of GE appears to be faster than in vertebrate, as the tissue dominant clustering of GE was only observed among closely related species, but not among species from different families (Figure S3). This pattern is consistent with a previous study which showed the overall GE of the three highly divergent species, including both monocots and dicots (diverge ~200 million years ago), are grouped according to species rather than organs [74]. In vertebrates, some tissues, such as brain, testis heart and muscle still demonstrates a strong tissue splicing signature, despite the dominant species-specific splicing background [20, 21]. Although all three tissues (root, leaves and flowers) used in our study did not show such tissue-specific splicing signature, some other plant tissues might. For example, the transcriptomes of sexual tissues have been shown to be substantially different from those of vegetative tissues, and the anthers harbor the most diverged specialized metabolome [74, 75]. Future studies that include transcriptome data of much more fine scaled tissue samples will provide new insights on this aspect.

We found that the AS resulted in transcripts with PTC that likely coupled with nonsense-mediated decay (NMD) for degradation, is more conserved than the other AS in plants (Figure 3A). Consistently, among six ultra-conserved AS events across different species including an ancient plant species (spreading earth moss), five produce PTC+ AS transcripts, indicating that PTC+ AS might be more important than it was previously thought. Previous studies also showed

that all human serine/arginine rich (SR) genes and some SR genes in plants produce PTC+ AS transcripts [17, 76, 77]. Furthermore, the junction regions that result in PTC+ AS transcripts in numerous splicing factors (SFs) are ultra-conserved between different kingdoms and the loss of the ancient PTC+ AS transcripts in paralogs though gene duplications would be quickly and repeatedly replaced by new raised distinct unproductive splicing [78]. Together, these results suggest that the unproductive splicing coupled with NMD can be viewed as a functional process that controls the abundance of active protein at a post-transcriptional level.

The distance between the 5'/3' nearest alternative splice sites (SS) and the authentic SS is the main determinant to distinguish AltD/AltA from constitutive splicing (Figure 4 and Table S1), respectively, among all five plant species. For a given spliced junction, the likelihood of AS decreases with the distance between the authentic and nearest alternative SS. Interestingly, similar pattern was also found in mammals, in which, the closer alternative SS to authentic SS were more likely used for AS [40, 59] than the more distant alternative SS, and the alternative SS within 6 nt to authentic SS was used in highest frequency [79]. The similar pattern found in both plants and mammals indicating that the mechanisms of generating AS, at least for AltD and AltA, are similar between these two kingdoms.

While the deep learning model for AltA achieved high precision and specificity among five species (AUC > 0.9), the models for AltD, although are better than by chance, performed less well than AltA (AUC < 0.8, Figure 4C and D, S13). This indicates that additional determinants that contribute to the regulations of AltD were not detected using our method. It is known that the mechanisms involved in AltD are more complex than AltA. For example, in both human and mouse, while both the presence and quantity of exon splicing enhancer (ESE) and exon splicing silencer (ESS) are known to be important for generating AltD [48], AltA is mainly

affected by the presence of the branch site (BS) and polypyrimidine tract (PPT) between authentic and alternative SS [80, 81]. In our attempts to identify ESE and ESS, although a few candidate sequence motifs that may act as putative splicing regulators were identified, none of them significantly contributed to the model prediction. Two non-exclusive possibilities may explain this. First, these motifs are not involved in splicing regulation process, although their density was significantly different between constitutively and alternatively spliced junctions. Second, they might be involved in tissue-specific regulations of AS, which is likely not contribute to the overall AltD prediction based on all three tissues. Future studies using different approaches, such as investigate the alteration of AS by introducing millions of random hexamers into specific regions of a gene junction in plant then measure the consequences of splicing, may allow us to detect more reliable splicing regulators of AltD in plants.

Although we found both junction size and SS for IR junctions are different from constitutively spliced junctions (Figure S10), the identified features could not resulted in an improved AS prediction from by chance, indicating some other key determinants for IR remain missing in the model. As the expression level of IR is usually low, the detection of which requires high sequencing depth (Figure S16). It is likely that the sequencing depth of the transcriptome data used in our study is not sufficient to detect all of the IR junctions, therefore, many IR junctions were not considered as IR in our dataset, which resulted in reduced prediction precision and power.

For both AltA and AltD, their rapid evolutions between closely related species are mainly due to the variations of the key sequence determinants near the SS (Figure 5, S14 and S15) and all of the key sequence determinants such as distance to alternative SS and *cis*-elements (BS, PPT, UA-rich tract for AltA) are all located within intronic region. Because intron sequences

diverge rapidly [82, 83], the process of which likely contributed to the rapid gain and loss of AS among different lineages and resulted in species-specific AS profile in plants. For example, a decreased distance between alternative SS and authentic SS can be resulted from a short deletion of intron sequence, which could results a gain of AS at the junction, and the consequence is likely to be shared among different tissues. Consistently, in vertebrates, the mutations that affect intronic splicing regulatory elements (SREs) were shown to be the main factor that resulted in the dominant species-specific splicing pattern [21]. However, our data do not exclude the possibility that the species-specific trans-factors, such as SR protein family, which have distinct numbers of homologues among species (Figure S17) [3, 18, 41], may have also contributed to the divergence of AS among different species [20, 84].

We also investigated some other factors, such as gene duplication (GD), DNA methylation and transposable element (TE) insertion, which were hypothesized to affect AS evolution [56, 57, 85]. However, except TE insertions, the effects of which were found to be species-specific, most factors did not show significantly effects on the levels of AS conservation between closely related species (Figure S15 and Table S2). The species-specific effects of TE on the AS conservation are largely due to different abundance of TE insertions in the genomes of different species [72, 73, 86, 87], suggesting genomic composition of each species might also affect the evolutionary alteration of AS.

Conclusions

We found AS profiles diverged rapidly in plant, which is largely due to rapid gain and loss of AS in each lineage, while a group of AS that generate PTC containing transcripts are ultra-conserved among highly distant plants. In addition, the alteration of a few key sequence

determinants of AltA and AltD, which are all located in the intron region, contributed to the fast divergence of AS between closely related plant species. These results provide new mechanistic insights on evolution of AS in plants and further shed lights on the role of post-transcriptional regulations in plant-environment interactions.

Materials and Methods

Read mapping, transcripts assembly and abundance estimation

The raw sequence reads were trimmed using AdapterRemoval (v1.1) [88] with parameters "--collapse --trimns --trimqualities 2 --minlength 36". The trimmed reads from each species were then aligned to the respective reference genome using Tophat2 (v2.0.6) [89], with maximum and minimum intron size set to 50,000 and 41 bp, respectively. The numbers of uniquely mapped reads and splice junctions (SJs) mapped reads were then counted using SAMtools (v0.1.19) [90] by searching "50" in the MAPQ string and "*N*" flag in the CIGAR string of the resulting BAM files. The uniquely mapped reads from each sample were sub-sampled into same sequencing-depth (16 millions) using SAMtools (v0.1.19) [90]. The mapping information and IDs of all download datasets deposited in Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) were listed in Table S3.

The transcripts of each species were assembled using Cufflinks (v2.2.0) [91] with the genome annotation as the reference. The open reading frame (ORF) of each transcript was analyzed using TransDecoder from TRINITY (v2.1.0) [92]. To estimate the expression level of genes/transcripts, all trimmed reads were re-mapped to the assembled transcripts using RSEM (v1.2.8) [93]. Transcripts per million (TPM) was calculated for each gene/transcript [94]. Only genes with TPM greater than five in at least one sample were considered as expressed gene.

AS detection

All AS analysis were based on splicing junctions obtained from the BAM files produced by Tophat2. To remove the false positive junctions that were likely due to non-specific or erroneous alignments, all original junctions were filtered based on overhang size greater than 13 bp, which was suggested in [3]. All filtered junctions were then used for AS identification and annotation using JUNCBASE v0.6 [95]. Due to the relatively low sequence depth of each individual sample of Brassicaceae RNA-seq data [68] (Table S4), we merged the BAM files of each three replicates together and random subsampled 17 million (the lowest depth among all merged samples) unique mapped reads from each merged to avoid the heterogeneity of sequencing depth.

The percent spliced index (PSI) of each AS event, which represents the relative ratio of two different isoforms generated by the AS was calculated in each sample. PSI = (number of reads to inclusion isoform) / (number of reads to inclusion isoform + number of reads to exclusion isoform) as suggested in [96].To avoid false-positive, PSI was calculated only for AS events that had a total read count equal or greater than ten.

Identification of conserved exon-exon junctions (EEJs) and AS

We separately extracted 100 bp sequence from the flanking upstream exon and downstream exon of each junction that have mapped read to support, and combine each side of exon sequence (in total 200 bp sequence) to represent the EEJ. The sequences of all EEJs from two species were search against with each other using TBLASTX (v.2.2.25) [97] to find homologues relationship (Figure S18). A python script was used to filter the TBLASTX results

based on the following requirements: (1) The gene pair contain the EEJs must be the orthologous gene pair between the two species; (2) the EEJ sequences between two species must be the best reciprocal blast hits based on the bit score; (3) the alignment position should start before 91 bp end and after 109 bp to ensure at least 3 amino acid (aa) from both the flanking upstream exon and downstream exon sequence, (4) alignment coverage >= 60 bp, (5) E-value <1E-3.

We only consider an AS event to be conserved if the same type of AS found on the conserved EEJs between two plant species.

Identification of AS that generate premature termination codons (PTC)

The junctions related to each AS event were mapped back to assembled transcripts, only AS which related junctions mapped to two unique transcripts (have no structure difference except the AS region) were kept to avoid the situation that the sequence differences of the two transcripts were caused by multiple AS events. Transcript was considered to have PTC if the stop codon of the longest ORF is at least 50 nucleotides upstream of an exon-exon boundary (the 50 nucleotides rule) [98]. The PTC generating AS events are defined as only one of the resulting transcripts contain PTC.

One-to-one orthologous identification and gene family size estimation

One-to-one orthologous gene pairs were predicted based on pair-wise sequence similarity between species of the corresponding dataset. First, we calculated the sequence similarities between all protein coding genes using BLASTP for the selected species and filtered the results based on E-value less than 1E-6. Second, we selected the groups of genes that represent the best reciprocal hits that are shared among all species from the corresponding dataset.

For calculating the gene family size, we first defined gene families among different species by using a similarity-based approach. To do so, we used all genes that were predicted from the respective genomes of each species. 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 gene families using the Markov cluster algorithm (mcl) [99]. The gene family size for a species is represented by the count of genes of this species within the corresponding gene family.

Correlation and clustering

For the pairwise comparison of AS, Spearman correlation and binary distance was applied to PSI ($0.05 < PSI < 0.95$ in at least one sample) and binary data (gene that has no AS in all four species are excluded), respectively. A non-parametric correlation was selected for PSI level because of its bimodal nature distribution (0 and 100). For the pairwise comparison of gene expression (GE), Pearson correlation was applied to $\log_2$ (TPM+1) of expressed genes to avoid infinite values.

The R package "pvcluster" was used for clustering of samples with 1,000 bootstrap replications. When we clustered and performed Principal-component analysis (PCA) of gene expression (GE), the TPM values were normalized by GC% (EDASeq package in R) and TMM (the trimmed mean of M-values).

Identification of possible alternative splice sites (SS) and regulatory sequences

The 5' and 3' splice site including 5 bp up and down-stream sequences of all EEJs were used as positive dataset, while the sequences extracted using the same method for all inter-GT (for 5' splice site) and inter-AG (for 3' splice site) within junction regions were used as background dataset. The putative SS motifs (6-mer) of both 5' and 3' SS were separately identified using Homer V3.12 [100] and only motifs present in at least 5% of total positive sequences and *P*-value<1E-20 were kept. The appearance of putative SS was identified using scanMotifGenomeWide, a perl script included in the Homer toolkits and only sequence regions with match score >2 were kept.

Homer was also used to identify the putative regulatory intronic and exonic sequence motifs (6-mer) of AltD, AltA and IR. The 50 bp up and down-stream sequence of 5' SS was regarded as exonic and intronic sequence and vice versa for 3' SS. For AltD and AltA, the related sequences of EEJs with AS were used as positive dataset, while 10,000 related sequences of EEJs without AS by random selection (due to the large number of sequences) were used as background dataset. The enriched motifs in the positive dataset were regarded as splicing enhancers, while the enriched motifs in the negative dataset were considered as splicing silencer. For IR, the related sequences from both splice sites of EEJs with IR were used as positive dataset and the same sequences from EEJs without IR were used as background dataset. The conserved motifs between species were identified using compareMotifs, a perl script included in the Homer toolkits and only one mismatch was allowed. To identify polypyrimidine (PPT), UA-rich tracts and branch site (BS) of each EEJ, we used perl scripts from [101, 102]. To estimate the effect of each putative sequence motif, PPT and UA-tracts, we calculated the AS frequency of EEJs containing or not containing the motif/tract [59]. Then for each motif/tract, the $\log_2$ odds ratio (effect size) with and without the motif/tract were calculated to quantify to what extent the

presence of the motif/tract increases or decreases the AS frequency compare to its absence:

$$Effect\ Size = \log 2 \frac{p(AS|motif)/(1-p(AS|motif))}{p(AS|-motif)/(1-p(AS|-motif))}$$

Decipher the splicing codes and AS conservation using deep learning algorithm

To investigate which sequence determinants contributed to the AS in plants, we construct multi-layer feed-forward artificial neural networks using H2O's deep learning algorithm ('h2o' package) in R 3.0.2 (R Development Core Team 2013). For each AS type, a matrix was created based on the information of all EEJs that contain the AS and other EEJs within the same gene. The AS status (either AS or constitutive) was considered as output and the features that were known to be associated with splicing recognition and regulation in eukaryotes [32, 59, 71] (listed in Table S1) were used as input for training the model. To reduce the background noise, we removed the EEJs which supported by less than five reads on average. In addition, because the number of constitutively spliced EEJs in all cases is much larger than alternative spliced EEJs, we randomly selected the same number of constitutive spliced EEJs as alternative spliced EEJs and combined them together with all alternative spliced EEJs as the full dataset (50% precision by chance). To train and test the deep neural networks (DNN), the full dataset was randomly split, which 60% of data were used for training, 20% used for validation and the other 20% held out for testing. We trained for a fixed number (10,000) of epochs or stopped the training once the top 10 model are within 1% of improvement, and selected the hyper-parameters that give the optimal AUC performance on the validation data. The model was then retrained using these selected hyper-parameters with the full dataset.

Using the similar approach, we constructed the model for AS conservation. For each AS type, a matrix was created based on the information of all orthologous EEJ pairs between two

species that contain the AS in at least one species. To reduce the background noise, any EEJ with multiple AS types, low number of support reads (less than five) or orthologous EEJ pair have different AS types were removed. The conservation levels (conserved, lost or gained in the other species) were used as the output of the model and the difference of features that were known to be important to AS and AS conservation [57-59, 103, 104] (listed in Table S2) between two species were used as input to train the model. Yass v1.15 [105] was used to align the SS' flanking sequences (combined 50 bp up-stream and down-stream sequences of 5'/3' splice site, 100 bp in total) of each orthologous EEJ pair, the similarity was calculated as: (length of alignment - number of gaps - number of mismatches) / (total sequence length). To reduce the bias from different transition types in the dataset (much higher proportion of loss/gain than conserved AS), the data used to train the model was selected as the ratio of 1:1:1 for conserved, lost and raised situation (33.3% precision by chance). Due to the rather small sample size of conserved AS, the model based on the same original data may be different as the randomly selected data of AS lost/raised were different each time. Therefore, the model construction process was repeated 10 time and the models that achieved the highest AUC for the complete dataset were considered.

Accession Numbers

The ID of all RNA-seq data that are deposited or downloaded from NCBI short reads archive (SRA) database for generating the results in this study are listed in Table S1.

List of abbreviations

AS: alternative splicing; IR: intron retention; AltA: alternative 3' accepter site; AltD: alternative 5' donor site; ES: exon skipping; PSI: percentage splicing index; SR: serine/arginine-rich; SJs: splicing junctions; bp: base pairs; ORF: open reading frame; SS: splice sites; GO: gene ontology; CDS: coding sequence; PTC: premature termination codons; EEJ: exon-exon junction; GD: gene duplication;

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Z.L. I.T.B. and S.X. designed the research. Z.L. T.B. and S.X. performed the experiments and analyzed the data. Z.L. and S.X. wrote the paper.

Acknowledgments

We thank Danell Seymour and Daniel Koenig for providing the methylation data, Michal Szczesniak for providing the in-house perl scripts. The work was supported for the funding by Swiss National Science Foundation (project number: PEBZP3-142886 to SX), the Marie Curie Intra-European Fellowship (IEF) (Project Number: 328935 to SX), the Max Planck Society, European Research Council advanced grant ClockworkGreen (Project number: 293926 to ITB).

Figures

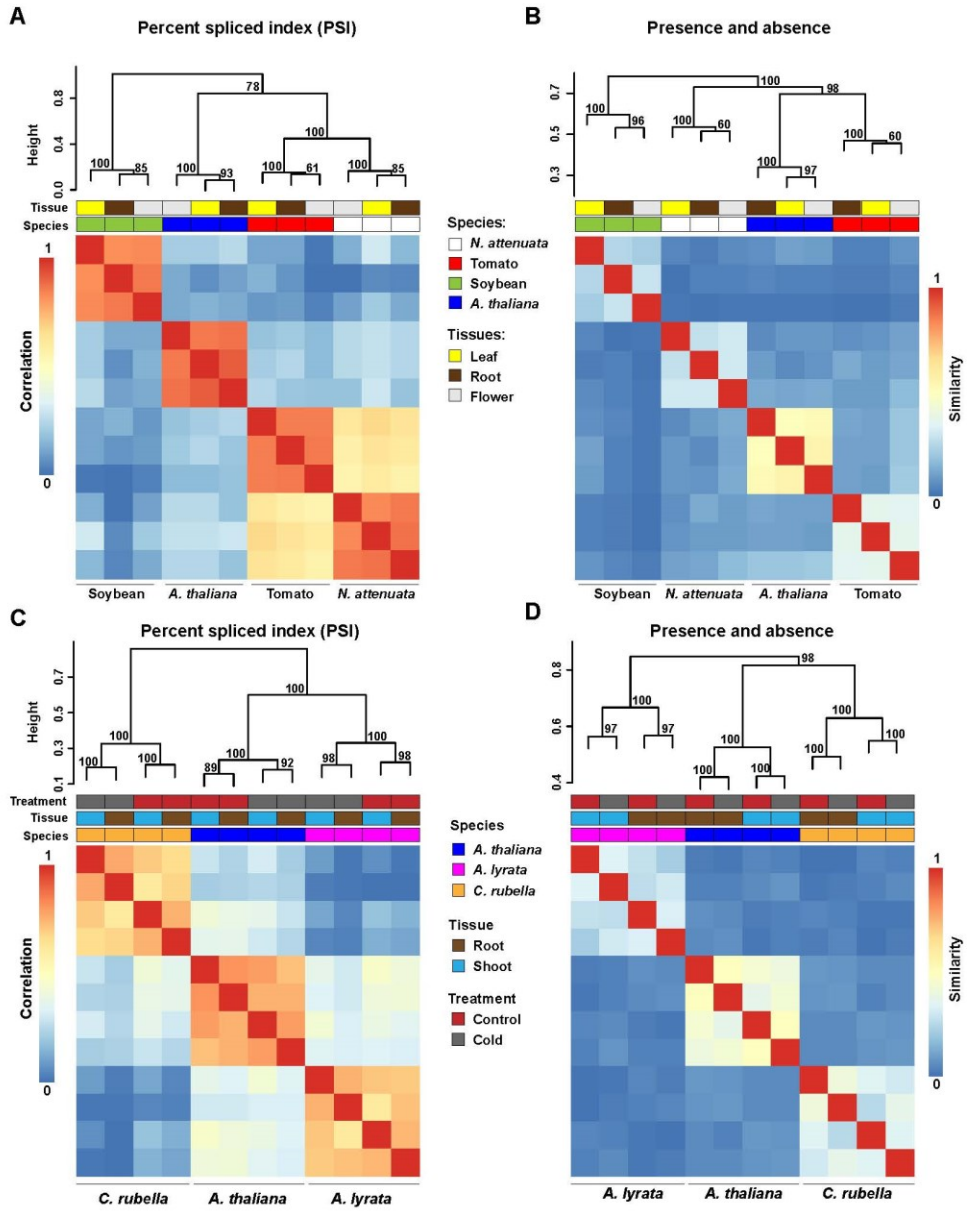


Figure 1. Species-specific clustering of alternative splicing (AS) among different plant species. A and C, heatmaps depict species-specific clustering based on PSI values among four

eudicots species (A) and three Brassicaceae species (C). The clustering are based on conserved splicing junctions (A and C: n = 502 and 5241, respectively). B and D, heatmaps depict species-specific clustering based on presence and absence of AS of the one-to-one orthologous genes. In total, junctions from 3857 (B) and 6262 (D) orthologous were used for the clustering. Numbers present in each branch node represent the approximately unbiased bootstrap value calculated from 1000 bootstrap replications. The color code above each heatmap represents species, tissue, and treatments.

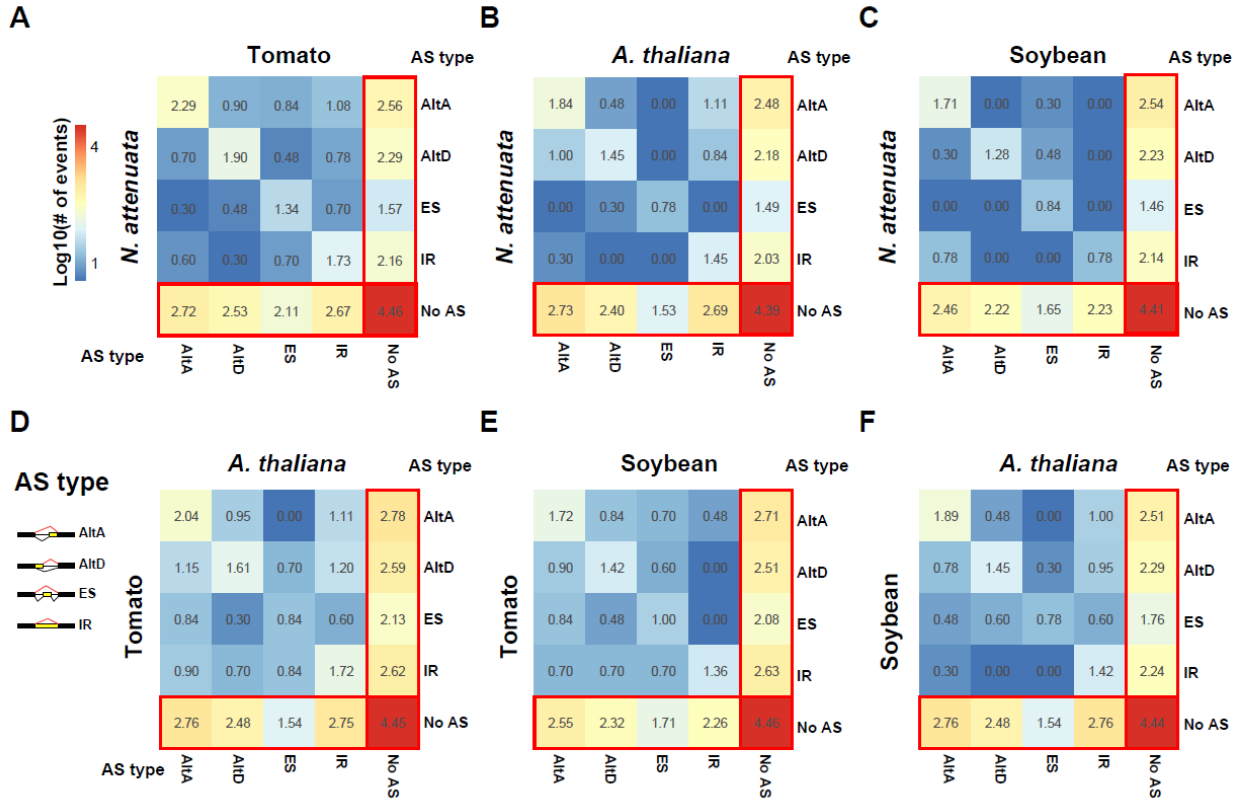


Figure 2. Transition spectrum of AS between each species pairs. (A) *N. attenuata* vs tomato, **(B)** *N. attenuata* vs *A. thaliana*, **(C)** *N. attenuata* vs soybean, **(D)** tomato vs *A. thaliana*, **(E)** tomato vs soybean and **(F)** soybean vs *A. thaliana*. The color of each grid refers to log10 transformed number of AS events. The transformed values are also shown in the middle of each grid. AltA: alternative 3' acceptor site; AltD: alternative 5' donor site; ES: exon skipping; IR: intron retention.

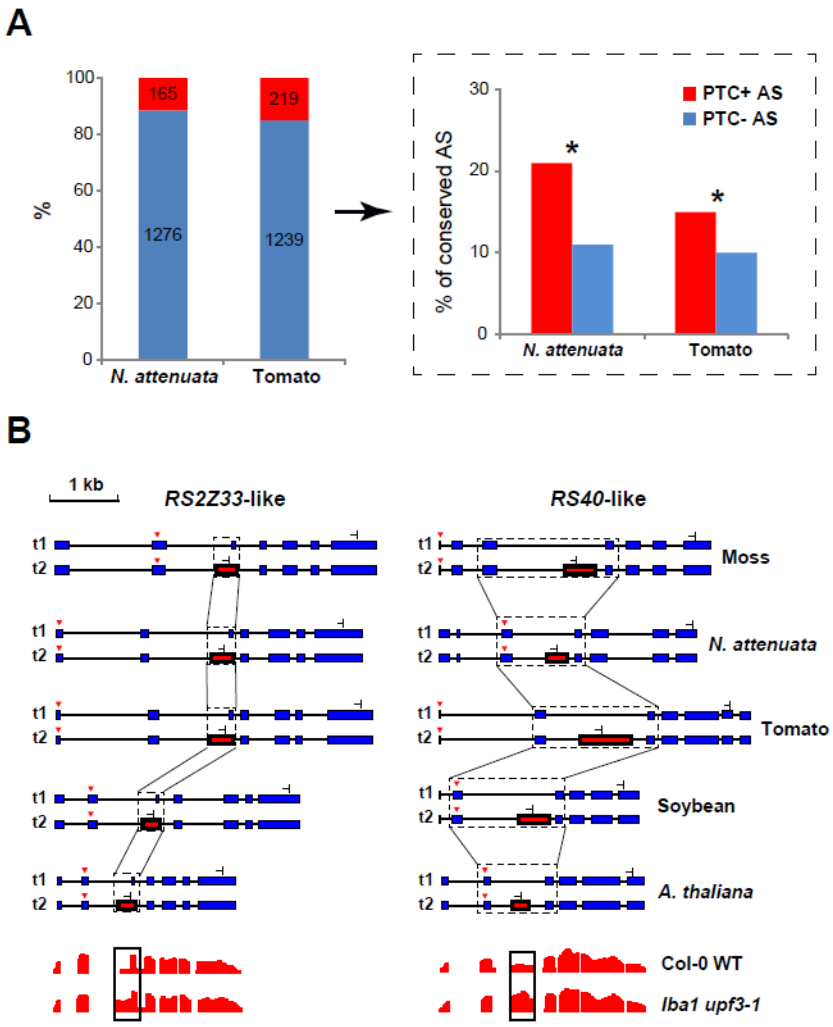


Figure 3. The conservation of alternative splicing between PTC+ and PTC AS. (A) The number and relative portions of PTC+/- AS in *N. attenuata* and tomato. The insert indicated by the black arrow depicts the likelihood of PTC+ and PTC- AS that are conserved between *N. attenuata* and Tomato. Asterisks indicate the significance as determined by Fisher's exact test ($P < 0.05$). **(B)** Conserved AS between moss and eudicots in Serine/arginine-rich splicing factor *RS2Z33*-like and *RS40*-like gene. The diagrams of the structure of transcripts generated by the AS in all five species, the dominant and minor transcripts are represented by t1 and t2, respectively. Constitutive exons are represented by blue box, alternative spliced exons are represented by red box and introns are represented by black solid lines. The black dotted boxes highlight the conserved AS and the start and stop codons are shown as red triangle and stop signs, respectively. The diagrams in the bot panel showed the relative read coverage of *AtRS2Z33* and *AtRS40* exons in wild type plant and *lba1 upf3-1* double mutants. The black box highlights the

coverage of the spliced region which is significantly increased in *lba1 upf3-1* double mutants (The diagrams are modified based on the data shown in http://gbrowse.cbio.mskcc.org/gb/gbrowse/NMD201)

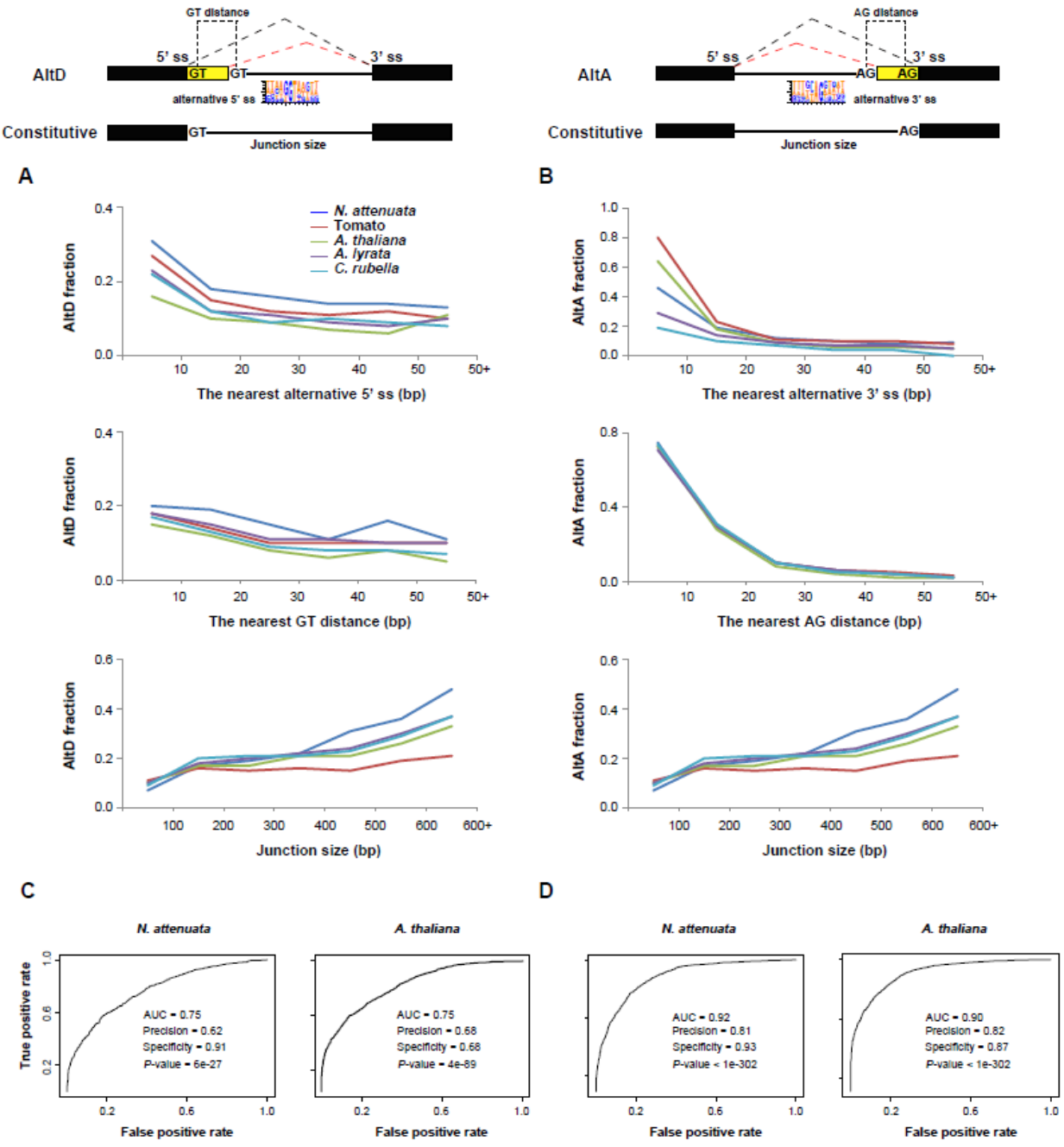


Figure 4. The determinants of alternative 5' donnor site (AltD) and alternative 3' acceptor site (AltA) in plants. (A) and (B), the frequencies of AltD/AltA on junctions with different distance between the authentic SS and nearest alternative SS (5' ss and 3' ss, respectively), and distance between authentic SS and the nearest inter GT/AG and junction size. (C) and (D), the area under the curve (AUC) plot of deep learning models using the key determinants of AltD and

AltA in *N. attenuata* and *A. thaliana*. The model performance including area under the curve (AUC), accuracy, specificity and significance are also shown.

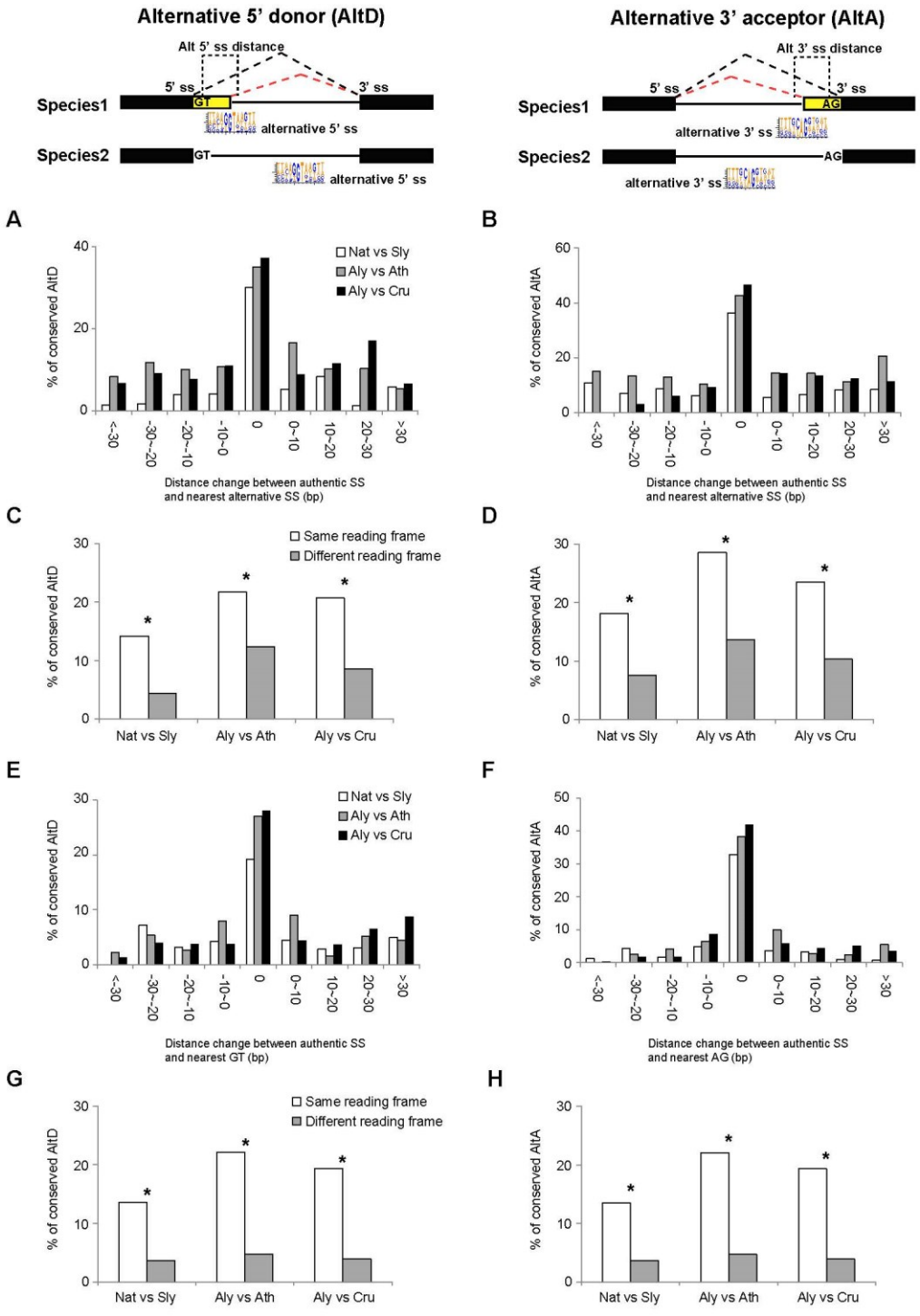


Figure 5. Features affect the conservation of AltD and AltA between closely related plant species. (A) and (B), the portion of conserved AltD/AltA decreases with changes of distance between authentic and alternative SS between two species. (C) and (D), the percent of conserved AltD/AltA in the group that the nearest alternative 5'/3' SS in the two species generate the same

or different ORF transcripts. (E) and (F), the portion of conserved AltD/AltA decreases with changes of distance between authentic SS and nearest inter-GT/AG sites between two species. (G) and (H) the percent of conserved AltD/AltA in the group that the nearest inter-GT/AG in the two species generate transcripts with same or different ORF (Nat: *N. attenuata*, Sly: Tomato, Ath: *A.thaliana*, Aly: *A. lyrata*). The asterisks indicate the significance as determined by Fisher's exact test ($P < 0.05$).

References

1. Staiger D, Brown JW: Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell* 2013, 25:3640-3656.
2. Mastrangelo AM, Marone D, Laido G, De Leonardis AM, De Vita P: Alternative splicing: enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci* 2012, 185-186:40-49.
3. Ling Z, Zhou W, Baldwin IT, Xu S: Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*. *The Plant journal : for cell and molecular biology* 2015, 84:228-243.
4. Howard BE, Hu Q, Babaoglu AC, Chandra M, Borghi M, Tan X, et al: High-throughput RNA sequencing of pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants. *PLoS One* 2013, 8:e74183.
5. Leviatan N, Alkan N, Leshkowitz D, Fluhr R: Genome-wide survey of cold stress regulated alternative splicing in *Arabidopsis thaliana* with tiling microarray. *PloS one* 2013, 8:e66511.
6. Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L: Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. *BMC Genomics* 2014, 15:431.
7. Rosloski SM, Singh A, Jali SS, Balasubramanian S, Weigel D, Grbic V: Functional analysis of splice variant expression of MADS AFFECTING FLOWERING 2 of Arabidopsis thaliana. *Plant molecular biology* 2013, 81:57-69.
8. Severing EI, van Dijk AD, Morabito G, Busscher-Lange J, Immink RG, van Ham RC: Predicting the impact of alternative splicing on plant MADS domain protein function. *PloS one* 2012, 7:e30524.
9. Barbazuk WB, Fu Y, McGinnis KM: Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome research* 2008, 18:1381-1392.
10. Reddy AS: Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol* 2007, 58:267-294.
11. Kazan K: Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged. *Trends Plant Sci* 2003, 8:468-471.
12. Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, et al: Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res* 2012, 40:2454-2469.
13. Chang YF, Imam JS, Wilkinson MF: The nonsense-mediated decay RNA surveillance pathway. *Annual review of biochemistry* 2007, 76:51-74.
14. Kervestin S, Jacobson A: NMD: a multifaceted response to premature translational termination. *Nature reviews Molecular cell biology* 2012, 13:700-712.
15. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, et al: Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & Development* 2007, 21:708-718.
16. Lareau LF, Brenner SE: Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Molecular Biology and Evolution* 2015, 32:1072-1079.
17. Kalyna M, Lopato S, Voronin V, Barta A: Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. *Nucleic Acids Res* 2006, 34:4395-4405.

18. Iida K, Go M: Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. *Mol Biol Evol* 2006, 23:1085-1094.
19. Darracq A, Adams KL: Features of evolutionarily conserved alternative splicing events between Brassica and Arabidopsis. *The New phytologist* 2013, 199:252-263.
20. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al: The evolutionary landscape of alternative splicing in vertebrate species. *Science* 2012, 338:1587-1593.
21. Merkin J, Russell C, Chen P, Burge CB: Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 2012, 338:1593-1599.
22. Severing EI, van Dijk ADJ, Stiekema WJ, van Ham RCHJ: Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *Bmc Genomics* 2009, 10.
23. Darracq A, Adams KL: Features of evolutionarily conserved alternative splicing events between *Brassica* and *Arabidopsis*. *New Phytologist* 2013, 199:252-263.
24. Satyawan D, Kim MY, Lee SH: Stochastic alternative splicing is prevalent in mungbean (*Vigna radiata*). *Plant biotechnology journal* 2016.
25. Streitner C, Koster T, Simpson CG, Shaw P, Danisman S, Brown JW, et al: An hnRNP-like RNA-binding protein affects alternative splicing by in vivo interaction with transcripts in Arabidopsis thaliana. *Nucleic acids research* 2012, 40:11240-11255.
26. Yang RL, Wang XF: Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *The Plant cell* 2013, 25:71-82.
27. Xu Q, Modrek B, Lee C: Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic acids research* 2002, 30:3754-3766.
28. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, et al: Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell* 2012, 46:884-892.
29. Black DL: Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 2003, 72:291-336.
30. Fu XD, Ares M, Jr.: Context-dependent control of alternative splicing by RNA-binding proteins. *Nature reviews Genetics* 2014, 15:689-701.
31. Simpson CG, Jennings SN, Clark GP, Thow G, Brown JW: Dual functionality of a plant U-rich intronic sequence element. *The Plant journal : for cell and molecular biology* 2004, 37:82-91.
32. Lewandowska D, Simpson CG, Clark GP, Jennings NS, Barciszewska-Pacak M, Lin CF, et al: Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell* 2004, 16:1340-1352.
33. Baek JM, Han P, Iandolino A, Cook DR: Characterization and comparison of intron structure and alternative splicing between *Medicago truncatula*, *Populus trichocarpa*, Arabidopsis and rice. *Plant molecular biology* 2008, 67:499-510.
34. Reddy AS, Shad Ali G: Plant serine/arginine-rich proteins: roles in precursor messenger RNA splicing, plant development, and stress responses. *Wiley interdisciplinary reviews RNA* 2011, 2:875-889.
35. Wang BB, Brendel V: The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome biology* 2004, 5:R102.

36. Gao H, Gordon-Kamm WJ, Lyznik LA: ASF/SF2-like maize pre-mRNA splicing factors affect splice site utilization and their transcripts are alternatively spliced. *Gene* 2004, 339:25-37.
37. Lopato S, Gattoni R, Fabini G, Stevenin J, Barta A: A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. *Plant molecular biology* 1999, 39:761-773.
38. Chen M, Manley JL: Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews Molecular cell biology* 2009, 10:741-754.
39. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, et al: HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008, 456:464-469.
40. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al: Deciphering the splicing code. *Nature* 2010, 465:53-59.
41. Isshiki M, Tsumoto A, Shimamoto K: The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. *The Plant cell* 2006, 18:146-158.
42. Richardson DN, Rogers MF, Labadorf A, Ben-Hur A, Guo H, Paterson AH, et al: Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into sub-family classification and extent of alternative splicing. *PloS one* 2011, 6:e24542.
43. Lorkovic ZJ: Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends in plant science* 2009, 14:229-236.
44. Pertea M, Mount SM, Salzberg SL: A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. *BMC bioinformatics* 2007, 8:159.
45. Thomas J, Palusa SG, Prasad KV, Ali GS, Surabhi GK, Ben-Hur A, et al: Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of SCL33 pre-mRNA. *The Plant journal : for cell and molecular biology* 2012, 72:935-946.
46. Yoshimura K, Yabuta Y, Ishikawa T, Shigeoka S: Identification of a cis element for tissue-specific alternative splicing of chloroplast ascorbate peroxidase pre-mRNA in higher plants. *J Biol Chem* 2002, 277:40623-40632.
47. Schoning JC, Streitner C, Meyer IM, Gao Y, Staiger D: Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis. *Nucleic acids research* 2008, 36:6977-6987.
48. Koren E, Lev-Maor G, Ast G: The emergence of alternative 3' and 5' splice site exons from constitutive exons. *PLoS computational biology* 2007, 3:e95.
49. Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, et al: The "alternative" choice of constitutive exons throughout evolution. *PLoS genetics* 2007, 3:e203.
50. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al: Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* 2014, 24:1774-1786.
51. Donahue CP, Muratore C, Wu JY, Kosik KS, Wolfe MS: Stabilization of the tau exon 10 stem loop alters pre-mRNA splicing. *The Journal of biological chemistry* 2006, 281:23302-23306.

52. Warf MB, Berglund JA: Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* 2010, 35:169-178.
53. Liu HX, Goodall GJ, Kole R, Filipowicz W: Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in *Nicotiana plumbaginifolia*. *The EMBO journal* 1995, 14:377-388.
54. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J: Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics* 2009, 41:376-381.
55. Schwartz S, Meshorer E, Ast G: Chromatin organization marks exon-intron structure. *Nature structural & molecular biology* 2009, 16:990-995.
56. Sorek R, Ast G, Graur D: Alu-containing exons are alternatively spliced. *Genome Res* 2002, 12:1060-1067.
57. Su Z, Wang J, Yu J, Huang X, Gu X: Evolution of alternative splicing after gene duplication. *Genome Res* 2006, 16:182-189.
58. Lambert MJ, Cochran WO, Wilde BM, Olsen KG, Cooper CD: Evidence for widespread subfunctionalization of splice forms in vertebrate genomes. *Genome Res* 2015, 25:624-632.
59. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G: Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* 2015, 163:698-711.
60. Reddy AS, Marquez Y, Kalyna M, Barta A: Complexity of the alternative splicing landscape in plants. *The Plant cell* 2013, 25:3657-3683.
61. Wang XT, Hu LJ, Wang XF, Li N, Xu CM, Gong L, et al: DNA methylation affects gene alternative splicing in plants: an example from rice. *Molecular Plant* 2016, 9:305-307.
62. Leung MK, Xiong HY, Lee LJ, Frey BJ: Deep learning of the tissue-regulated splicing code. *Bioinformatics* 2014, 30:i121-129.
63. Mamoshina P, Vieira A, Putin E, Zhavoronkov A: Applications of deep learning in biomedicine. *Molecular pharmaceutics* 2016, 13:1445-1454.
64. Alipanahi B, Delong A, Weirauch MT, Frey BJ: Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 2015, 33:831-+.
65. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M: Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res* 2012, 22:1184-1195.
66. Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, et al: Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* 2014, 26:996-1008.
67. Aoki K, Yano K, Suzuki A, Kawamura S, Sakurai N, Suda K, et al: Large-scale analysis of full-length cDNAs from the tomato (Solanum lycopersicum) cultivar Micro-Tom, a reference system for the Solanaceae genomics. *BMC genomics* 2010, 11:210.
68. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D: Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet* 2014, 10:e1004785.
69. Li W, Lin WD, Ray P, Lan P, Schmidt W: Genome-wide detection of condition-sensitive alternative splicing in Arabidopsis roots. *Plant Physiol* 2013, 162:1750-1763.

70. Drechsel G, Kahles A, Kesarwani AK, Stauffer E, Behr J, Drewe P, et al: Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *The Plant cell* 2013, 25:3726-3742.
71. Kandul NP, Noor MA: Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. *BMC genetics* 2009, 10:67.
72. Tomato Genome C: The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 2012, 485:635-641.
73. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al: The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* 2011, 43:476-481.
74. Yang R, Wang X: Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *The Plant cell* 2013, 25:71-82.
75. Li D, Heiling S, Baldwin IT, Gaquerel E: Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. *Proceedings of the National Academy of Sciences of the United States of America* 2016.
76. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE: Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 2007, 446:926-929.
77. Palusa SG, Reddy AS: Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay. *The New phytologist* 2010, 185:83-89.
78. Lareau LF, Brenner SE: Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Molecular biology and evolution* 2015, 32:1072-1079.
79. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ: Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *Rna* 2006, 12:2047-2056.
80. Smith CW, Porro EB, Patton JG, Nadal-Ginard B: Scanning from an independently specified branch point defines the 3' splice site of mammalian introns. *Nature* 1989, 342:243-247.
81. Smith CW, Chu TT, Nadal-Ginard B: Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Molecular and cellular biology* 1993, 13:4939-4952.
82. Hare MP, Palumbi SR: High intron sequence conservation across three mammalian orders suggests functional constraints. *Molecular biology and evolution* 2003, 20:969-978.
83. Mattick JS: Introns: evolution and function. *Current opinion in genetics & development* 1994, 4:823-831.
84. Ast G: How did alternative splicing evolve? *Nature reviews Genetics* 2004, 5:773-782.
85. Flores K, Wolschin F, Corneveaux JJ, Allen AN, Huentelman MJ, Amdam GV: Genome-wide association between DNA methylation and alternative splicing in an invertebrate. *BMC genomics* 2012, 13:480.
86. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al: The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature genetics* 2013, 45:831-835.

87. Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, et al: The tobacco genome sequence and its comparison with those of tomato and potato. *Nature communications* 2014, 5:3833.
88. Lindgreen S: AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes* 2012, 5:337.
89. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25:1105-1111.
90. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009, 25:2078-2079.
91. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 2012, 7:562-578.
92. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 2011, 29:644-652.
93. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 2011, 12:323.
94. Wagner GP, Kin K, Lynch VJ: Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften* 2012, 131:281-285.
95. Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, et al: Conservation of an RNA regulatory map between Drosophila and mammals. *Genome research* 2011, 21:193-202.
96. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, et al: The developmental transcriptome of Drosophila melanogaster. *Nature* 2011, 471:473-479.
97. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *Journal of molecular biology* 1990, 215:403-410.
98. Nagy E, Maquat LE: A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends in biochemical sciences* 1998, 23:198-199.
99. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* 2002, 30:1575-1584.
100. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* 2010, 38:576-589.
101. Szczesniak MW, Kabza M, Pokrzywa R, Gudys A, Makalowska I: ERISdb: a database of plant splice sites and splicing signals. *Plant Cell Physiol* 2013, 54:e10.
102. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G: Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome research* 2008, 18:88-103.
103. Kelley DR, Hendrickson DG, Tenen D, Rinn JL: Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol* 2014, 15:537.
104. Li Q, Xiao G, Zhu YX: Single-nucleotide resolution mapping of the Gossypium raimondii transcriptome reveals a new mechanism for alternative splicing of introns. *Molecular plant* 2014, 7:829-840.

105. Noe L, Kucherov G: YASS: enhancing the sensitivity of DNA similarity search. *Nucleic acids research* 2005, 33:W540-543.

Supplemental Figures and Tables

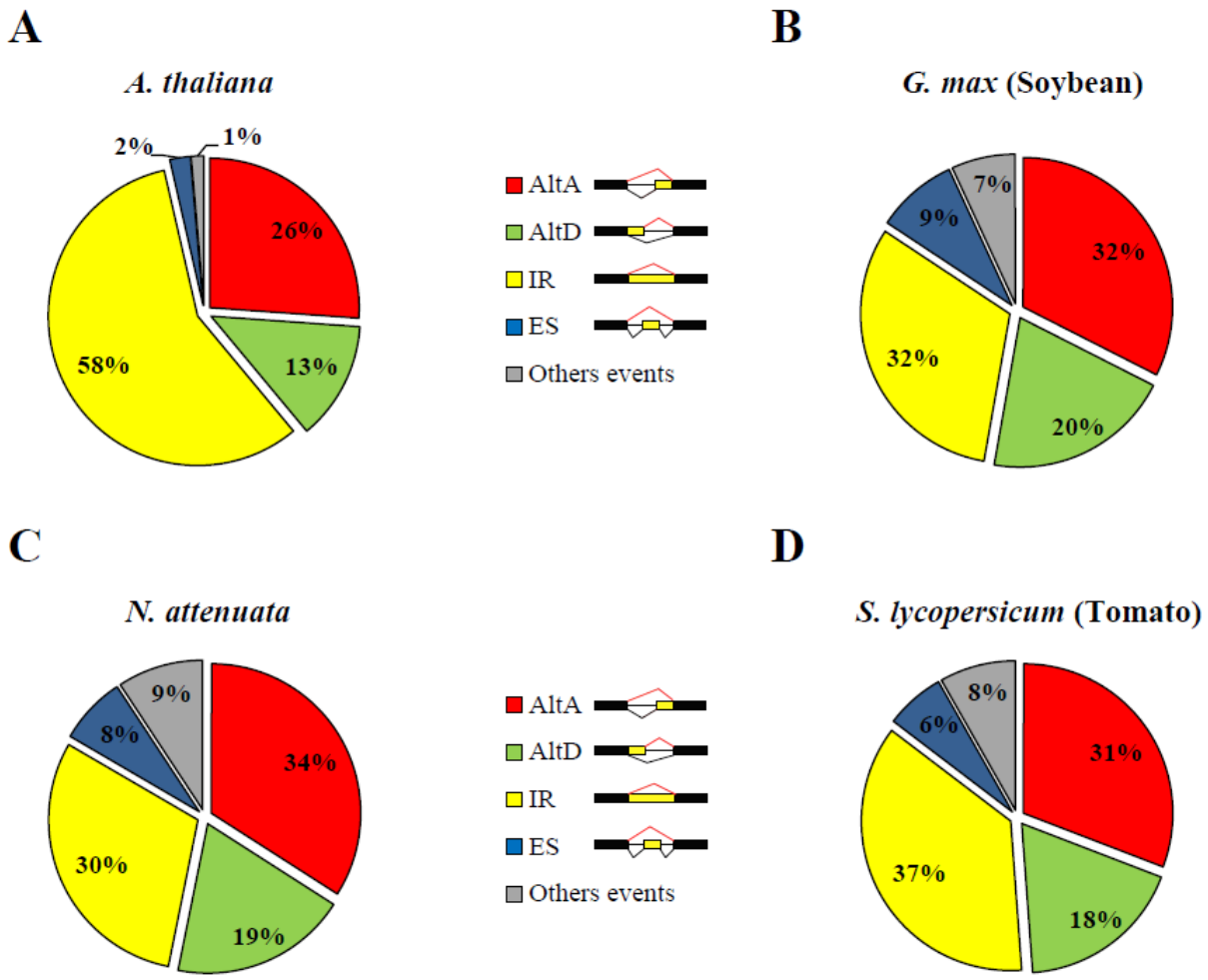


Figure S1. The distribution of different types of alternative splicing (AS) events in (A) *A. thaliana*, **(B)** *G. max*, **(C)** *S. lycopersicum*, **(D)** *N. attenuata* (IR: intron retention; AltA: alternative 3' acceptor site; AltD: alternative 5' donor site; ES: exon skipping.).

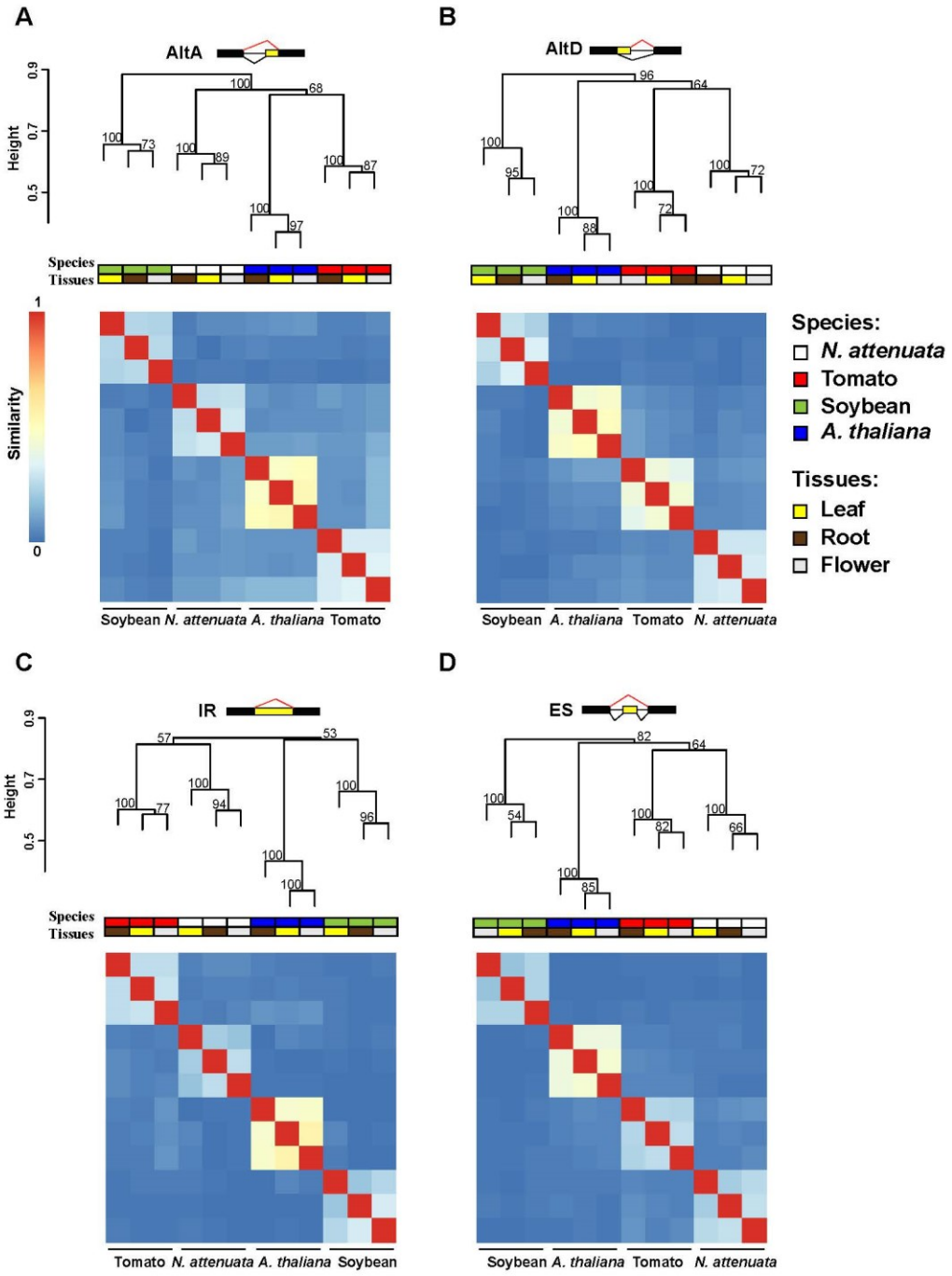


Figure S2. Species-specific clustering of alternative splicing (AS) among different plant species. Heatmaps depict species-specific clustering based on presence and absence of AS among one-to-one orthologous genes in all four species (n = 3857) for **(A)** alternative 3' acceptor site (AltA), **(B)** alternative 5' donor site (AltD), **(C)** intron retention (IR) and **(D)** exon skipping

(ES). Numbers at each branch node represent the approximately unbiased bootstrap value calculated from 1000 bootstrap replications. The color code of the heatmaps represents species, tissue, and treatments.

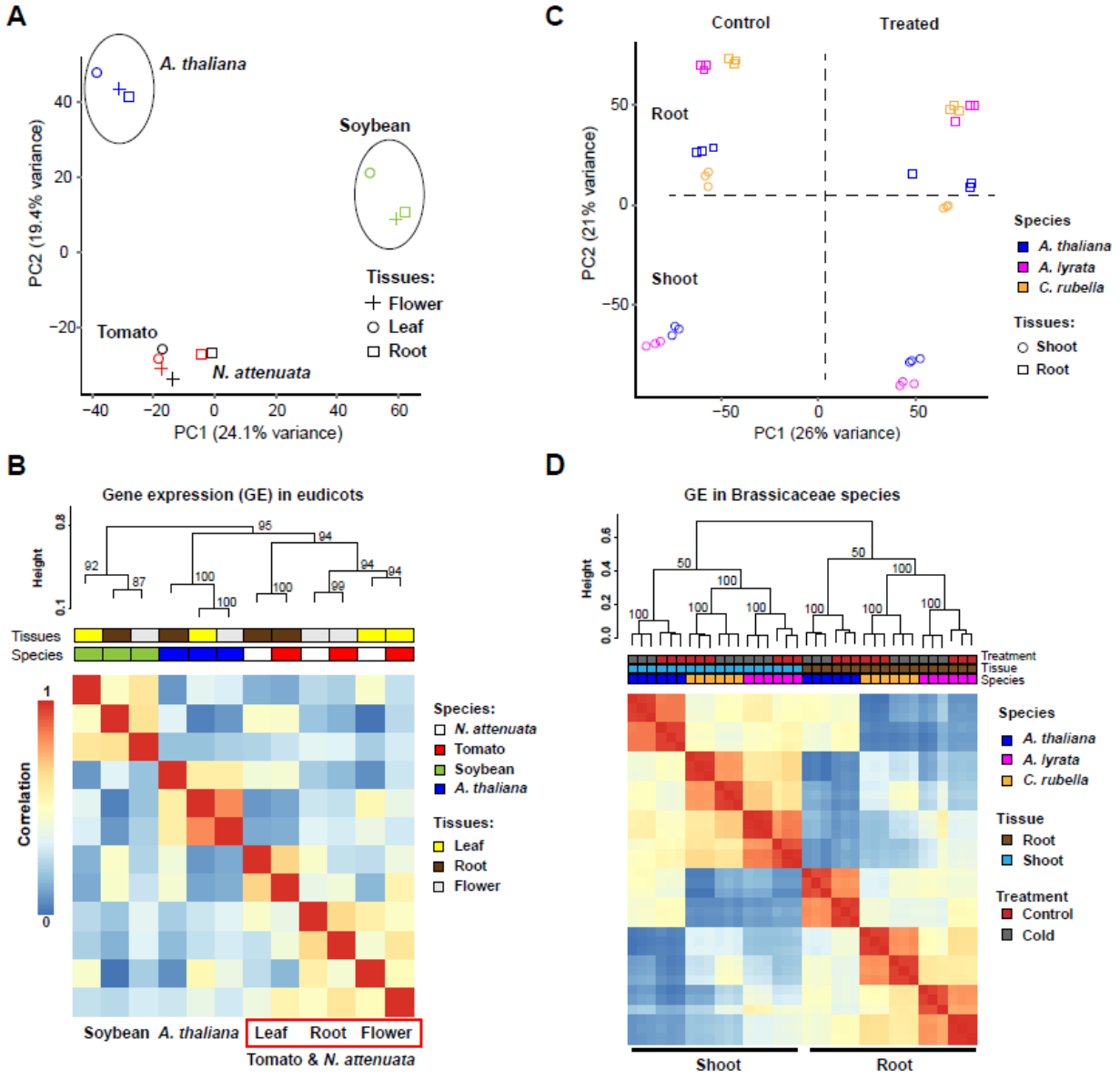


Figure S3. Conservation of gene expression (GE) in eudicots. (A) and (B) The PCA analysis and the clustering of different samples based on normalized expression values of one-to-one orthologous genes present in all four compared eudicots (n = 5745). **(C) and (D)** The PCA analysis and the clustering of different samples based on normalized expression values of one-to-one orthologous present in all three compared Brassicaceae species (n = 15969). For (B) and (D), the complete linkage hierarchical clustering was used with the distance measured by Pearson correlations. Numbers at each branch node represent the approximately unbiased bootstrap value calculated from 1000 bootstrap replications.

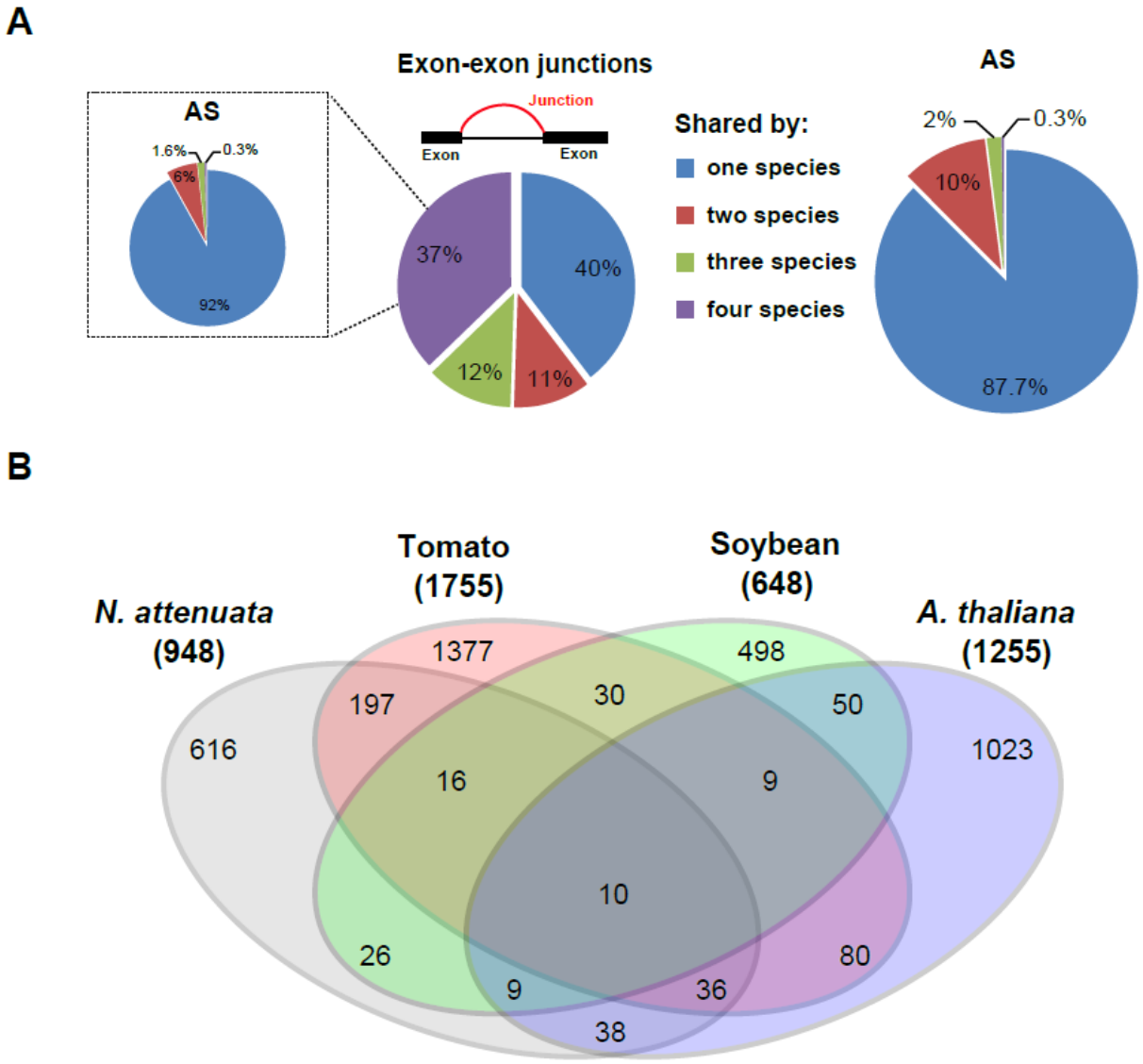


Figure S4. Comparative profiling of conserved junctions and alternative splicing in eudicots. (A) The proportions of exon-exon junctions (EEJs) and AS that are conserved among different number of species. The insert panel depicts the conservation of AS events identified on the most conserved junctions (shared by all four species). **(B)** Venn diagram depicts the distribution of 4,015 AS events that were conserved in at least two species.

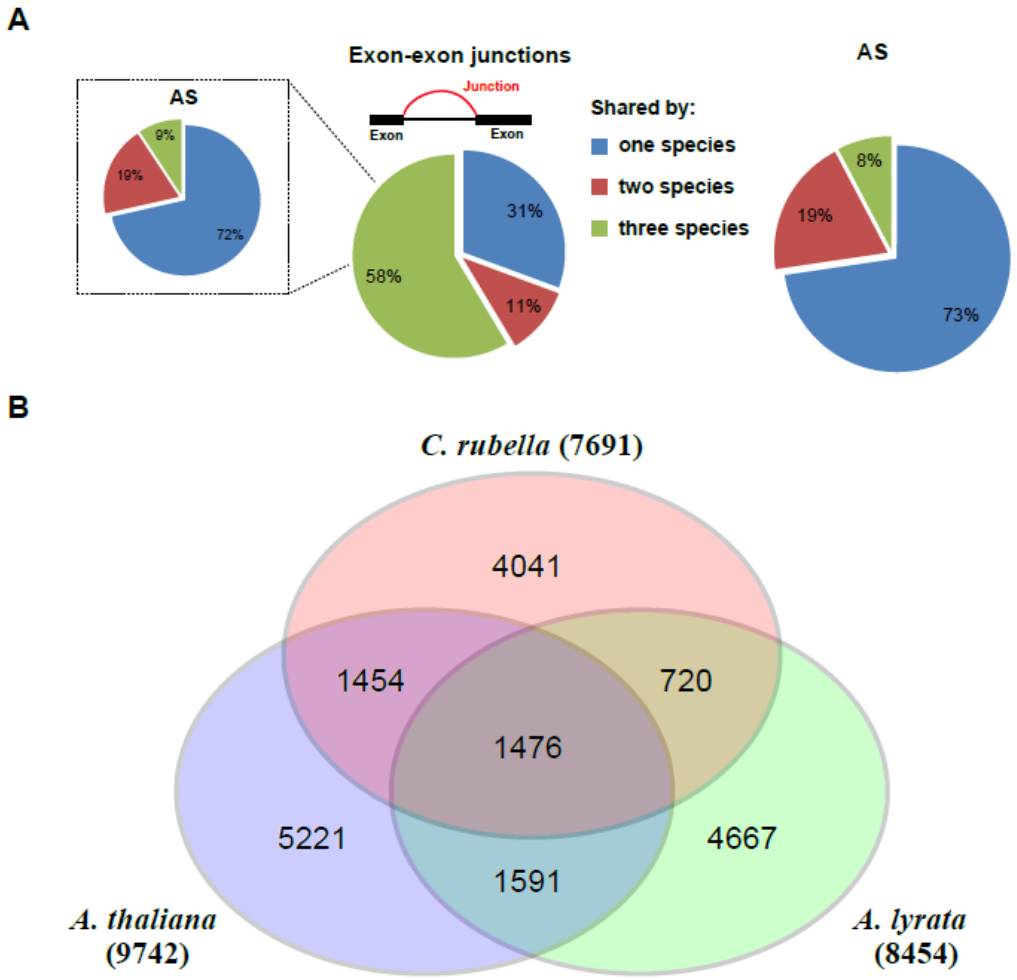


Figure S5. Comparative profiling of conserved junction and alternative splicing (within one-to-one orthrologs) in three Brassicaceae species. (A) The proportion of exon-exon junctions (EEJs) and AS that are conserved among different number of species. The insert panel depicts the conservation of AS events identified on the most conserved junctions (shared by all three species). **(B)** Venn diagram depicts the distribution of 19,170 AS events that were conserved in at least two species.

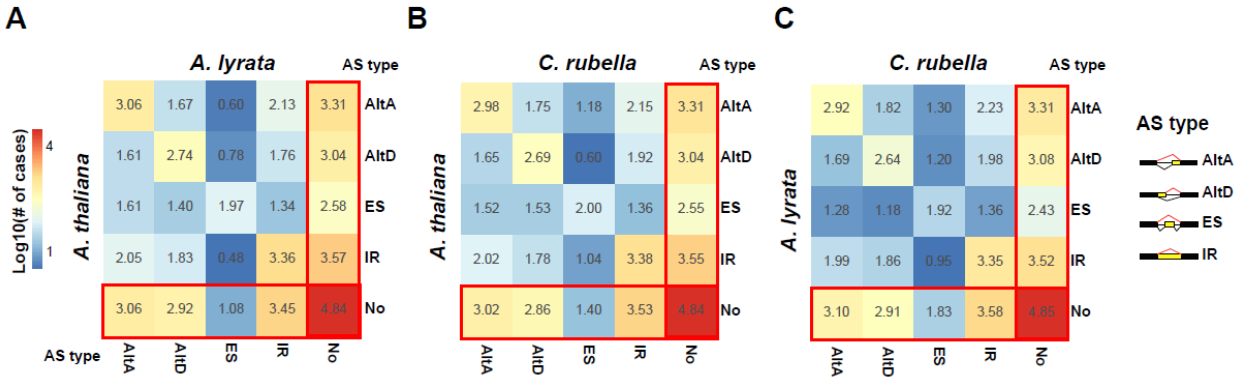


Figure S6. The transition spectrum among different types of alternative splicing (AS) between species pairs. (A) *A. thaliana* vs *A. lyrata*, **(B)** *A. thaliana* vs *C. rubella*, **(C)** *A. lyrata* vs *C. rubella* (AltA: alternative 3' acceptor site; AltD: alternative 5' donor site; ES: exon skipping; IR: intron retention).

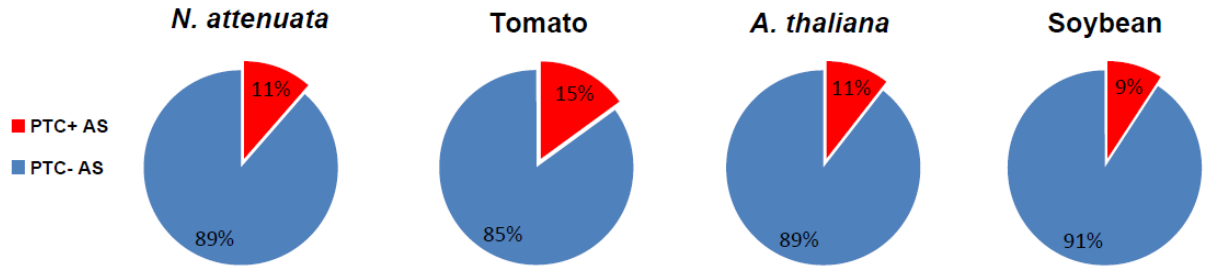


Figure S7. The propotion of alternative splicing (AS) that generate PTC (PTC+) or not (PTC-) in *N. attenuata*, tomato, *A. thaliana* and soybean.

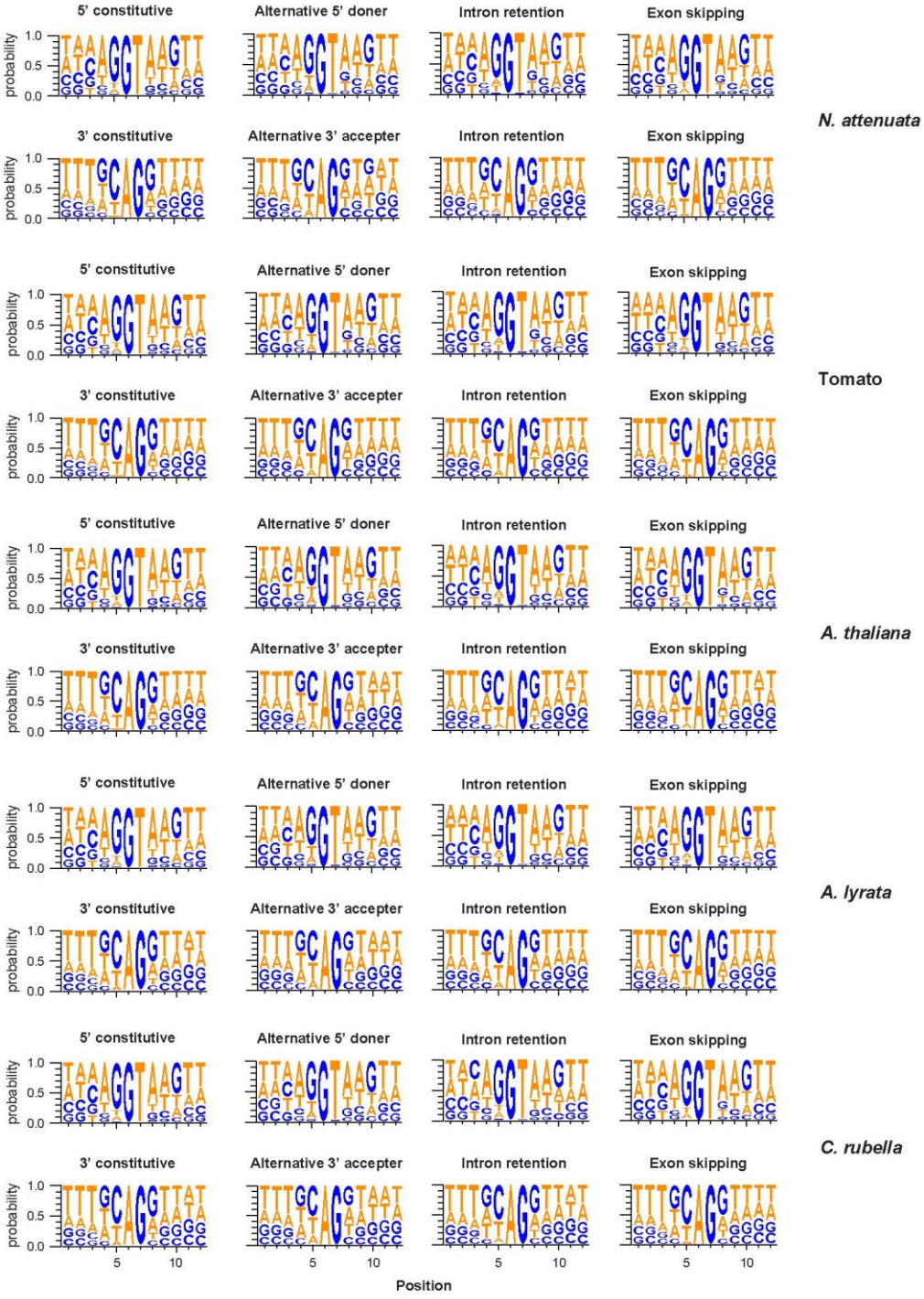


Figure S8. The probability of DNA bases surrounding splice sites with different AS types compared to regular splice sites in five plant species.

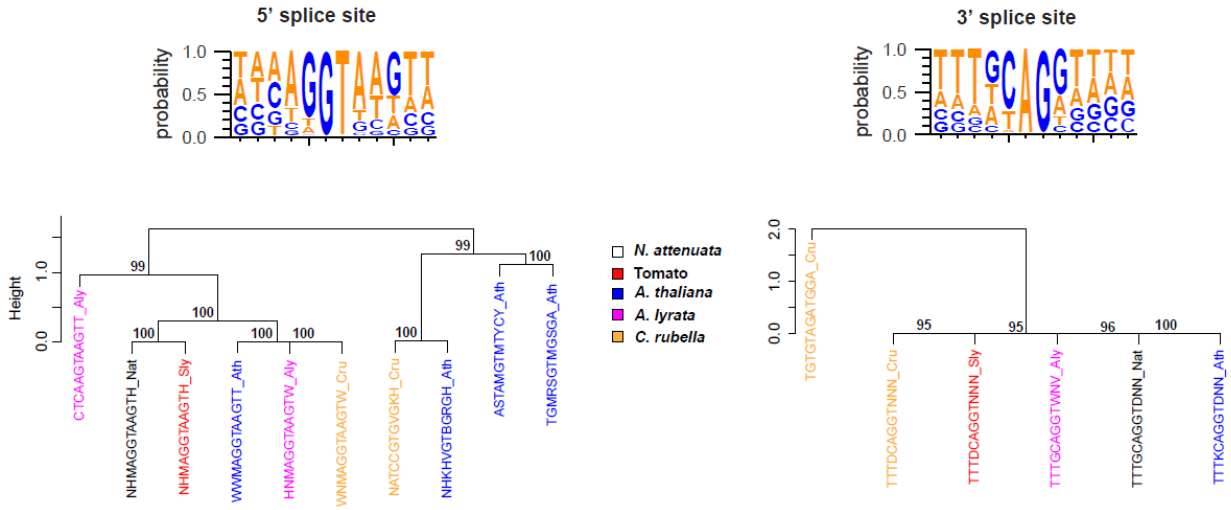


Figure S9. The complete linkage hierarchical clustering of splice site (SS) motifs among different plant species. The distance was measured by Pearson correlation. Number at each branch node represents the approximately unbiased bootstrap value calculated from 1000 bootstrap replications.

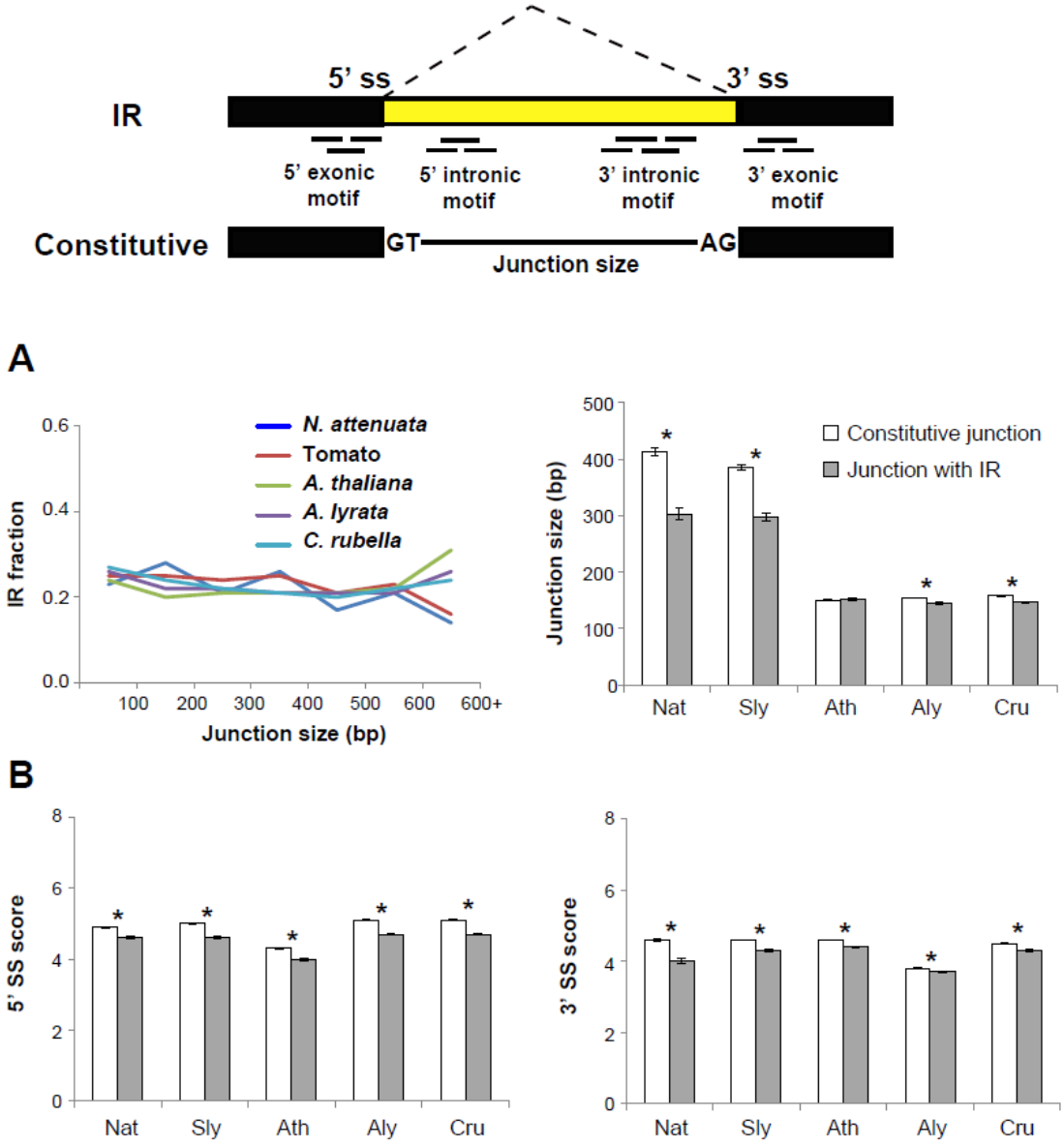


Figure S10. The determinants of intron retention (IR) in plants. (A), the frequencies of IR for junctions with different size and the average size between constitutive junction and junction with IR. (B), the average 5' and 3' splice site (SS) score between constitutive junction and junction with IR. The asterisk indicates a significant difference determined by Student's *t*-test ($P < 0.05$) and the error bars depict standard error (SE).

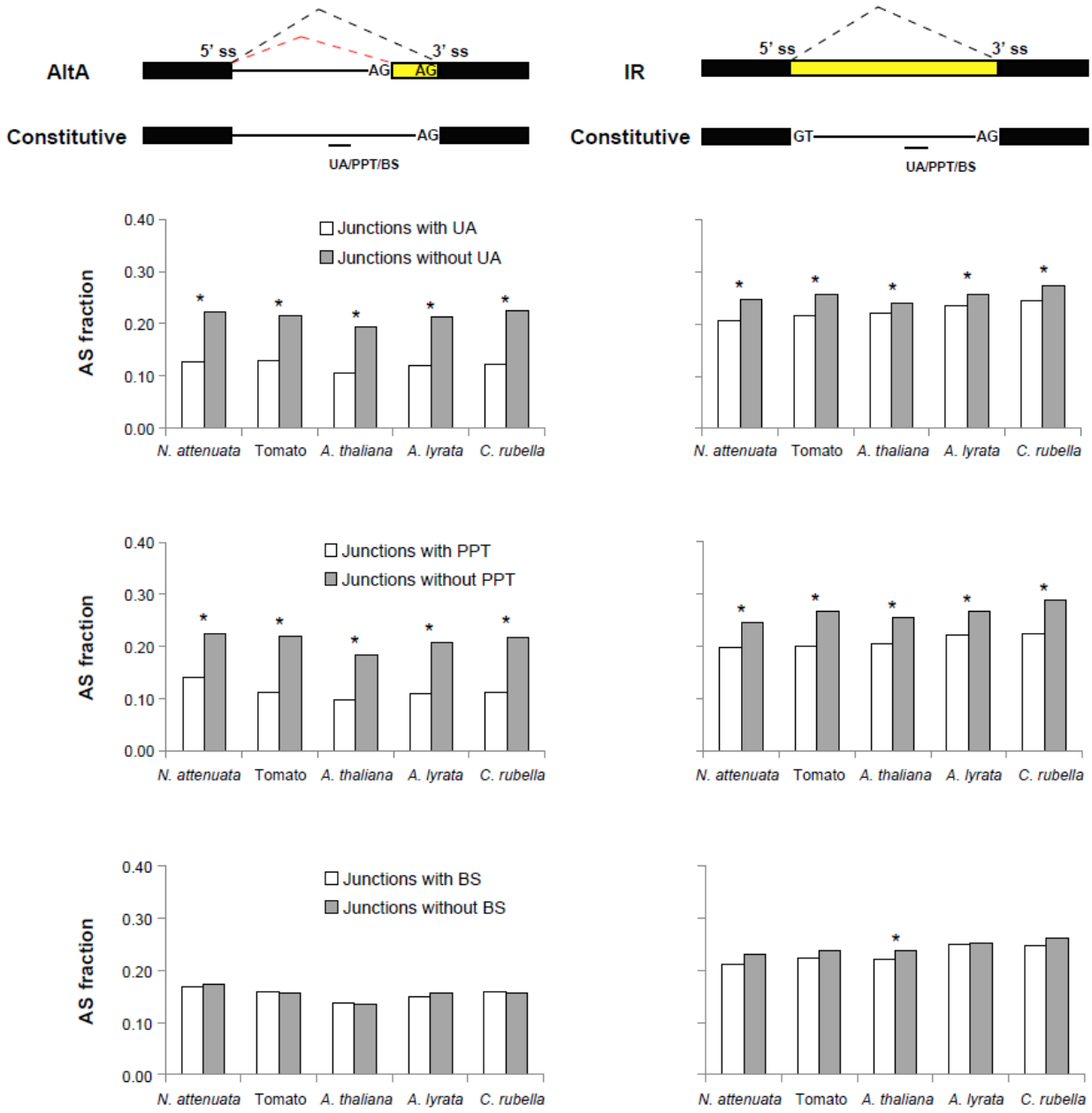


Figure S11. The effect of UA-rich, polypyrimidine tract (PPT) and branch site (BS) on alternative acceptor (AltA) and intron retention (IR) in plants. The frequencies of AltA and IR between junctions with and without the sequence tracts are shown. The asterisks indicate the significance as determined by Fisher's exact test ($P < 0.05$).

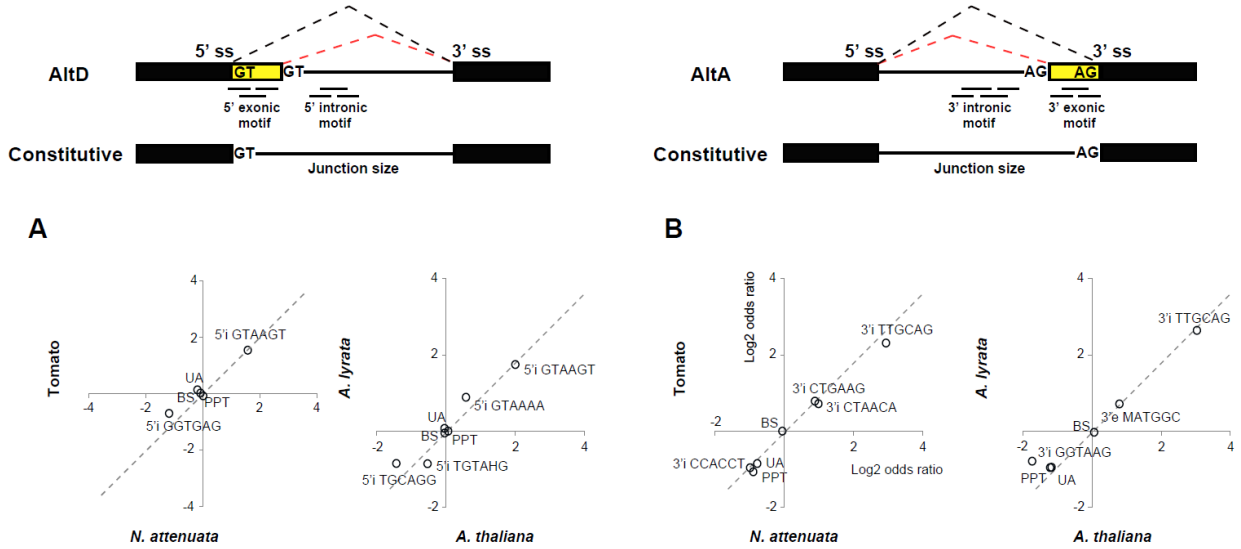


Figure S12. The effect size of conserved 6-mer motifs and features identified in (A) alternative 5' donnor site (AltD) and (B) alternative 3' acceptor site (AltA) between species pairs in Solanaceae and Brassicaceae.

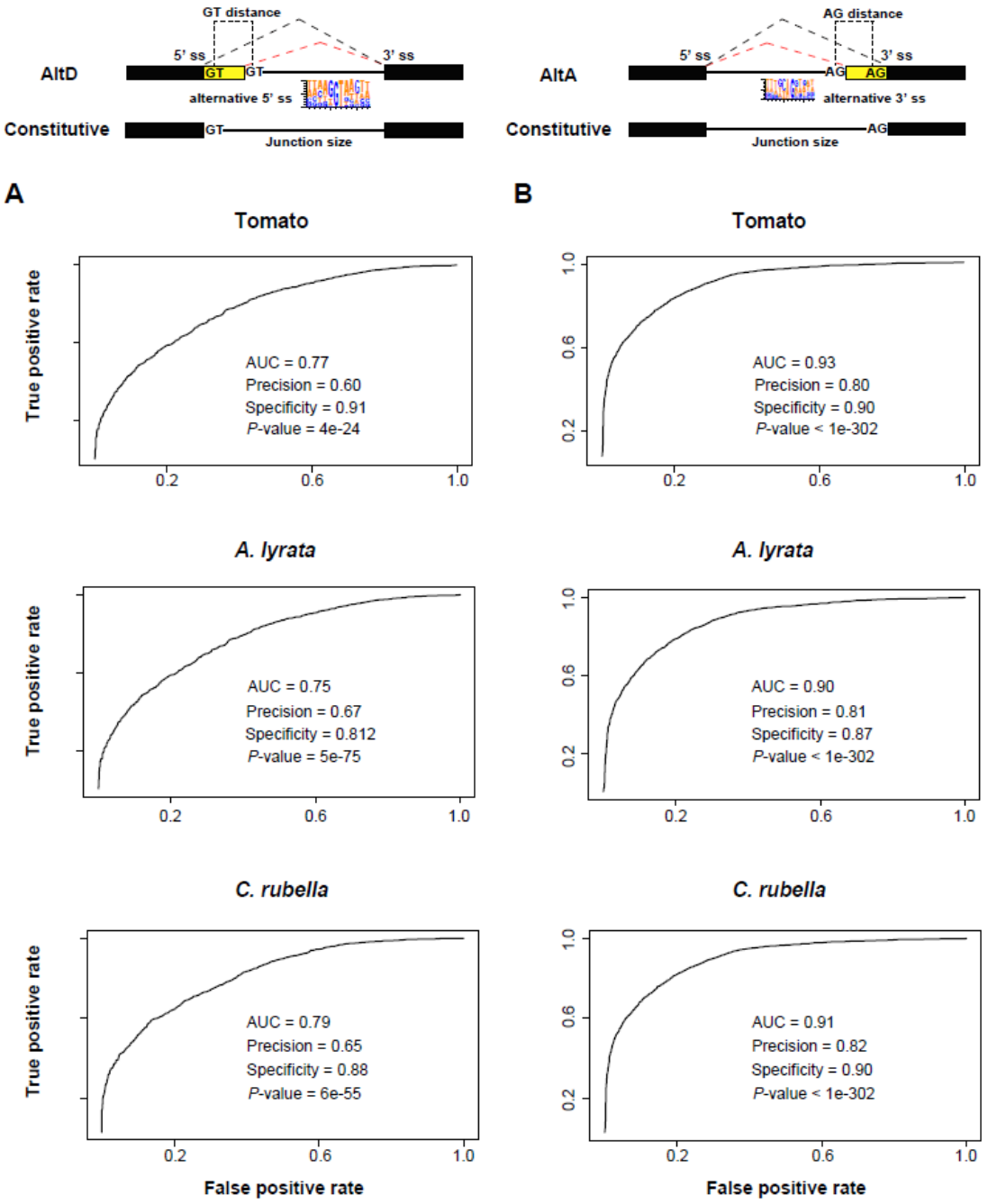


Figure S13. The area under the curve (AUC) of deep learning models using different key features of (A) alternative 5' donor (AltD) and (B) alternative 3' acceptor (AltA) in tomato, *A. lyrata* and *C. rubella*. For each model, the model performance including area under the curve (AUC), accuracy, specificity and significance are also shown.

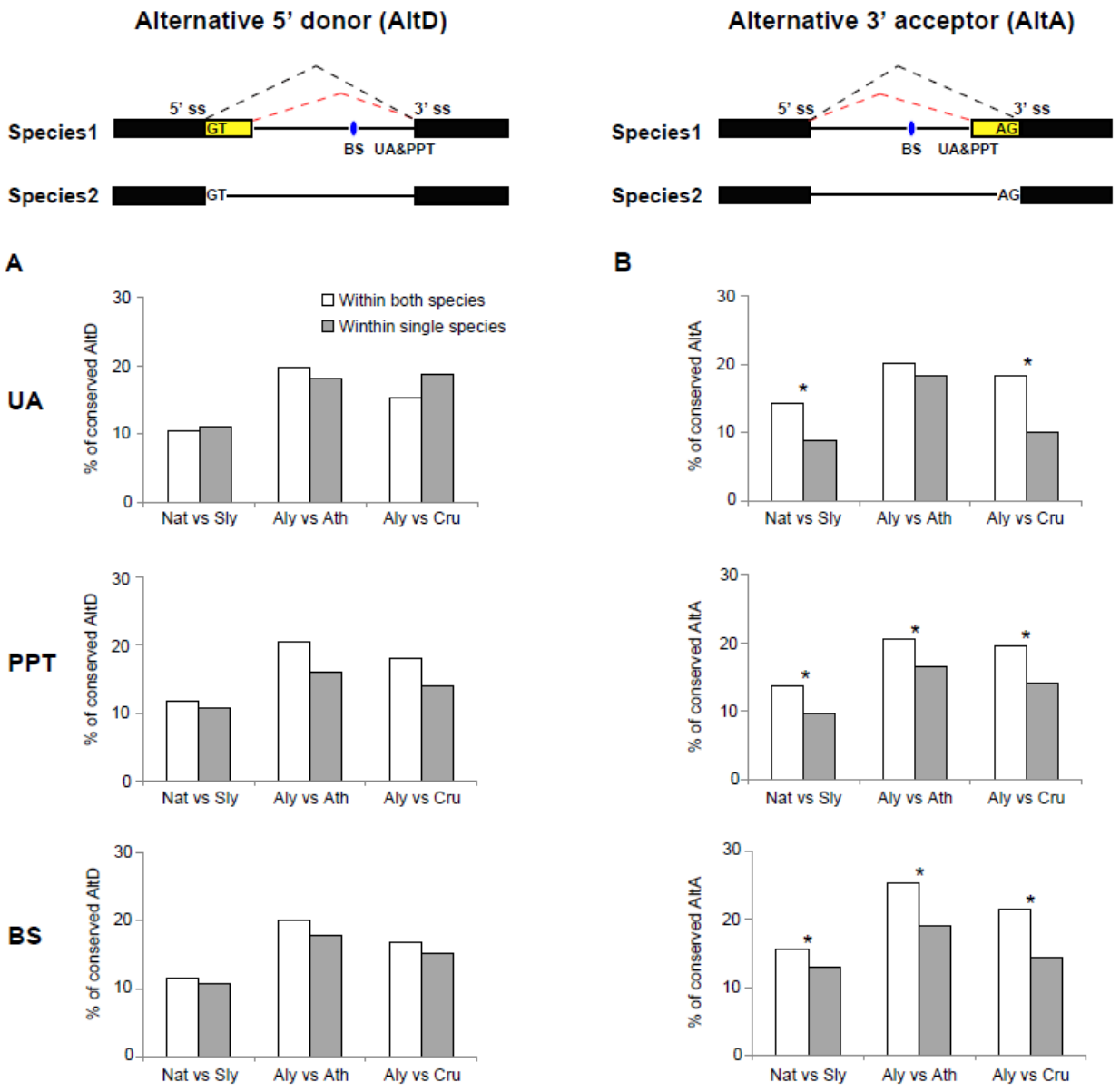


Figure S14. The differences of the *cis*-regulatory elements affect the turnover rate of (A) AltD and (B) AltA between plant species. The bar-plots display the percentage of conserved AS in the group that UA, PPT and BS are different or same between two closely related species (Nat: *N. attenuata*, Sly: Tomato, Ath: *A.thaliana*, Aly: *A. lyrata*), the asterisks indicate the significance as determined by Fisher's exact test ($P < 0.05$).

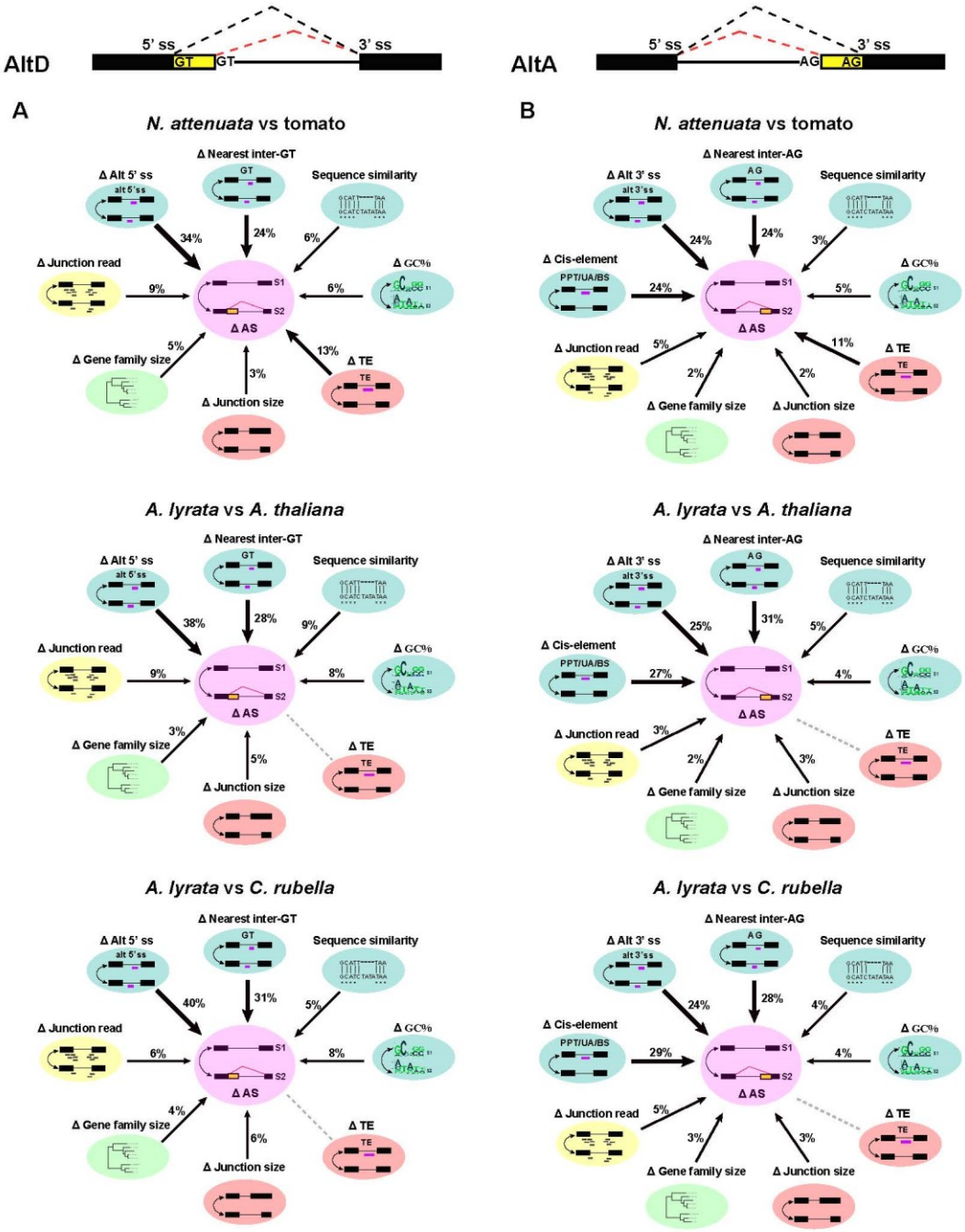


Figure S15. Factors that affect the rapid turnover of AS between plant species. Factors involved in AS turnover between *N. attenuata* and tomato, *A. lyrata* and *A. thaliana* and *A. lyrata* and *C. rubella* of (A) alternative 5' donnor site (AltD) and (B) alternative 3' acceptor site

(AltA) are shown. The number upon each arrow indicates the proportion of each factor that contribute to the model, the thickness of arrows are used to scale the contribution. Factors with no contribution are indicated by grey dotted lines.

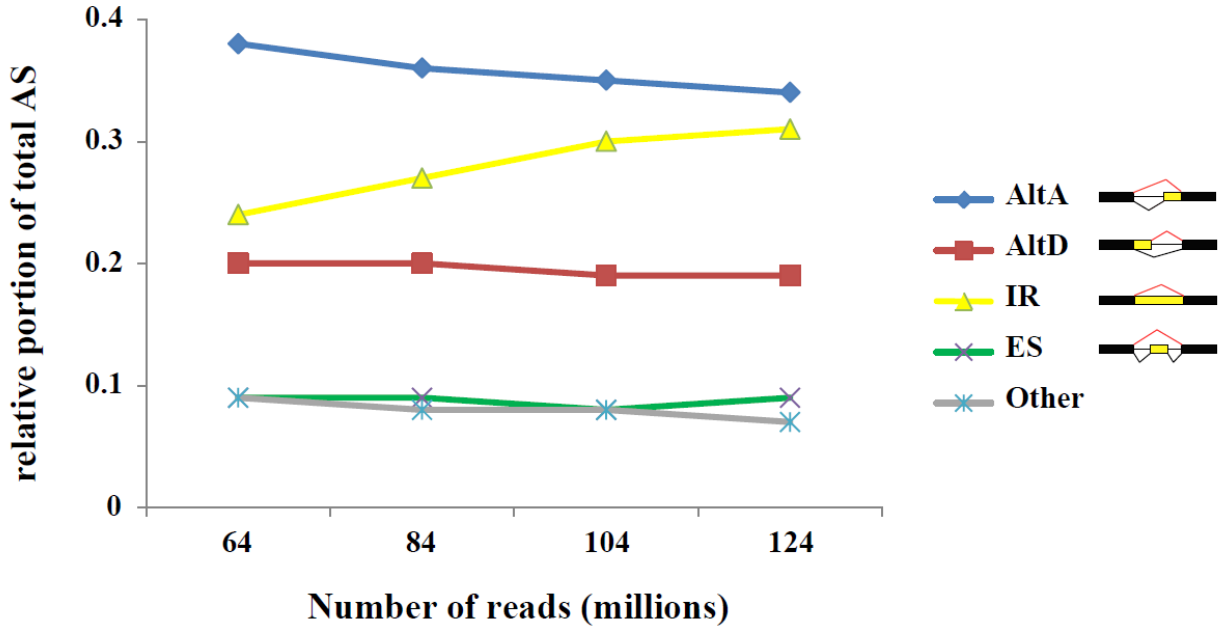


Figure S16. The composition of AS types detected in leaf and root samples of *N. attenuata* with different sequencing depth. (AltA: 3' accepter site (AltA); AltD: alternative 5' donor site; IR: intron retention; ES: exon skipping).

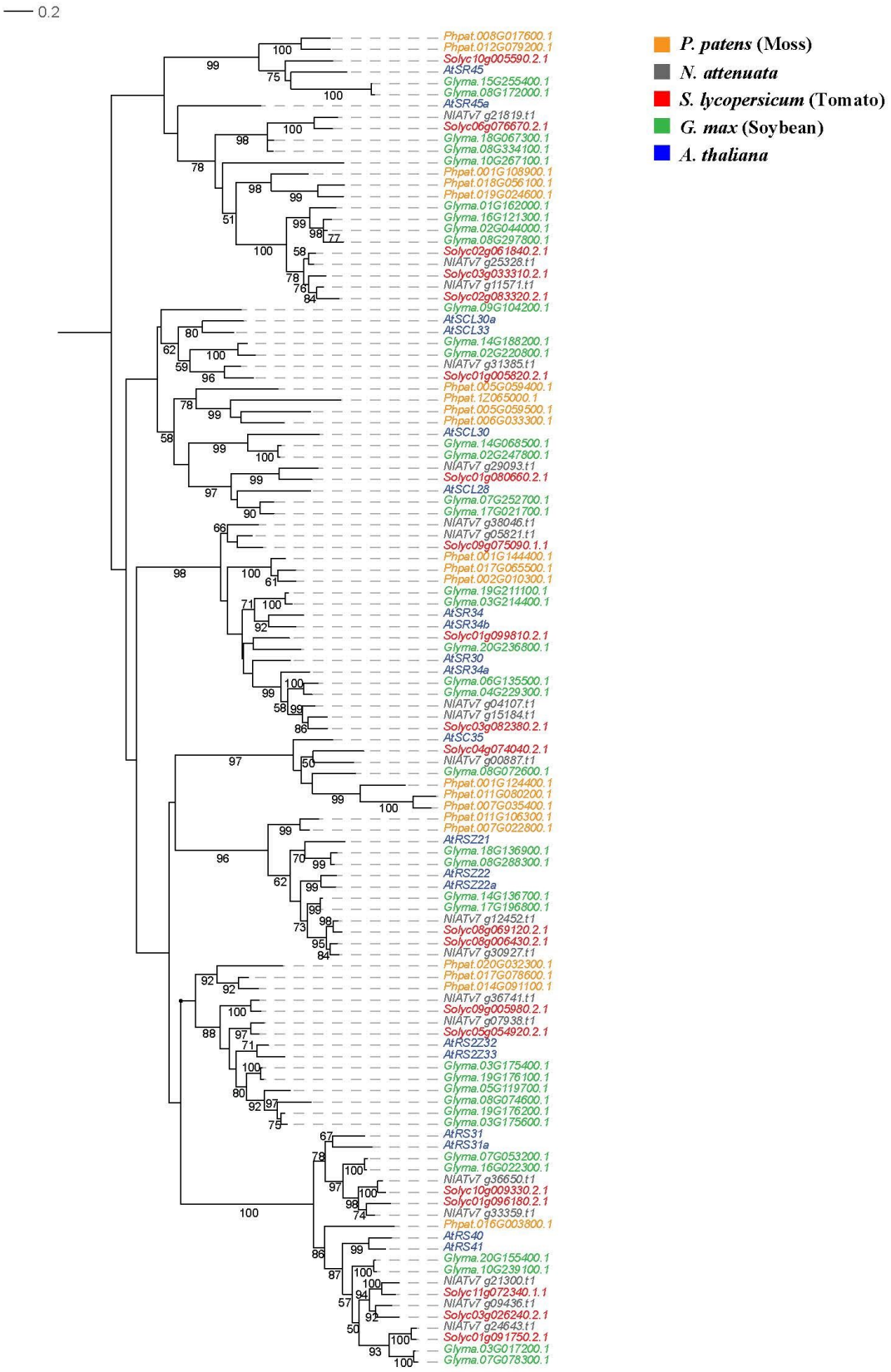

0.2
P. patens (Moss)
N. attenuata
S. lycopersicum (Tomato)
G. max (Soybean)
A. thaliana
Phpat.008G017600.1
Phpat.012G079200.1
Solyc10g005590.2.1
AtSR45
Glyma.15G255400.1
Glyma.08G172000.1
AtSR45a
NIATv7 g21819.t1
Solyc06g076670.2.1
Glyma.18G067300.1
Glyma.08G334100.1
Glyma.10G267100.1
Phpat.001G108900.1
Phpat.018G056100.1
Phpat.019G024600.1
Glyma.01G162000.1
Glyma.16G121300.1
Glyma.02G044000.1
Glyma.08G297800.1
Solyc02g061840.2.1
NIATv7 g25328.t1
Solyc03g033910.2.1
NIATv7 g11571.t1
Solyc02g083320.2.1
Glyma.09G104200.1
AtSCL30a
AtSCL33
Glyma.14G188200.1
Glyma.02G220800.1
NIATv7 g31385.t1
Solyc01g005820.2.1
Phpat.005G059400.1
Phpat.1Z065000.1
Phpat.005G059500.1
Phpat.006G033300.1
AtSCL30
Glyma.14G068500.1
Glyma.02G247800.1
NIATv7 g29093.t1
Solyc01g080660.2.1
AtSCL28
Glyma.07G252700.1
Glyma.17G021700.1
NIATv7 g38046.t1
NIATv7 g05821.t1
Solyc09g075090.1.1
Phpat.001G144400.1
Phpat.017G065500.1
Phpat.002G010300.1
Glyma.19G211100.1
Glyma.03G214400.1
AtSR34
AtSR34b
Solyc01g099810.2.1
Glyma.20G236800.1
AtSR30
AtSR34a
Glyma.06G135500.1
Glyma.04G229300.1
NIATv7 g04107.t1
NIATv7 g15184.t1
Solyc03g082380.2.1
AtSC35
Solyc04g074040.2.1
NIATv7 g00887.t1
Glyma.08G072600.1
Phpat.001G124400.1
Phpat.011G080200.1
Phpat.007G035400.1
Phpat.011G106300.1
Phpat.007G022800.1
AtRSZ21
Glyma.18G136900.1
Glyma.08G288300.1
AtRSZ22
AtRSZ22a
Glyma.14G136700.1
Glyma.17G196800.1
NIATv7 g12452.t1
Solyc08g069120.2.1
Solyc08g006430.2.1
NIATv7 g30927.t1
Phpat.020G032300.1
Phpat.017G078600.1
Phpat.014G091100.1
NIATv7 g36741.t1
Solyc09g005980.2.1
NIATv7 g07938.t1
Solyc05g054920.2.1
AtRS2232
AtRS2233
Glyma.03G175400.1
Glyma.19G176100.1
Glyma.05G119700.1
Glyma.08G074600.1
Glyma.19G176200.1
Glyma.03G175600.1
AtRS31
AtRS31a
Glyma.07G053200.1
Glyma.16G022300.1
NIATv7 g36650.t1
Solyc10g009330.2.1
Solyc01g096180.2.1
NIATv7 g33359.t1
Phpat.016G003800.1
AtRS40
AtRS41
Glyma.20G155400.1
Glyma.10G239100.1
NIATv7 g21300.t1
Solyc11g072340.1.1
NIATv7 g09436.t1
Solyc03g026240.2.1
NIATv7 g24643.t1
Solyc01g091750.2.1
Glyma.03G017200.1
Glyma.07G078300.1

Figure S17. Phylogenetic tree of SR and SR-like genes in moss and four eudicots. The tree was constructed using a maximum likelihood method. Different colors represent different species and only bootstrap values greater than 50 are shown above each branch.

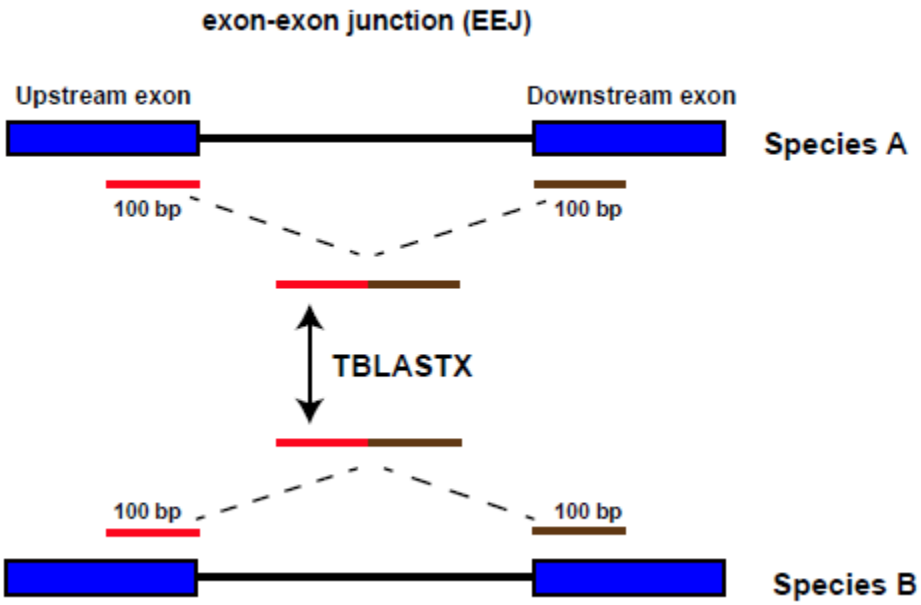


Figure S18. The diagram showing the process of identifying conserved exon-exon junctions (EEJs) between two species.

Supplemental Tables

Table S1: the relative importance of each factor that contribute to the splicing model of AltD and AltA (separate excel sheet).

Table S2: the relative importance of each factor that contribute to the AS conervation model of AltD and AltA (separate excel sheet).

Table S3. RNA-seq coverage and alignment statistics of the four eudicots

Species	Reference genome	Tissues	# trimmed reads	# uniquely mapped reads	% uniquely mapped reads	# reads uniquely mapped to junctions	% reads uniquely mapped to junctions	Read length (bp)	Read type	Sequence Read Archive (SRA) run ID
N. attenuata	NIATTr2	Root	41,318,960	37,387,919	90.49	9,483,825	22.95	100	Paired-end	SRR1951027
		Leaf	41,849,942	37,128,217	88.72	9,230,047	22.06	100	Paired-end	SRR1950961
		Flower	18,550,688	17,598,129	94.87	4,546,878	24.51	100	Paired-end	NA
S. lycopersicum	Slycopersicum_225	Root	26,603,412	23,573,422	88.61	5,909,485	22.21	90	Paired-end	SRR1210486
		Leaf	21,421,650	20,306,633	94.79	3,291,892	15.37	50	Paired-end	SRR570060
		Flower	24,779,060	23,369,504	94.31	6,847,338	27.63	100	Paired-end	SRR988531
G. max	Gmax_275_v2.0	Root	63,547,192	51,894,450	81.66	13,333,131	20.98	100	Paired-end	SRR1174205
		Leaf	62,592,462	54,526,801	87.11	15,090,482	24.11	100	Paired-end	SRR1174227
		Flower	88,313,226	75,241,464	85.20	17,024,663	19.28	100	Paired-end	SRR1174220
A. thaliana	Athaliana_167	Root	201,682,843	188,521,244	93.47	61,907,600	30.70	100	Single end	SRR505743
		Leaf	206,859,837	192,213,103	92.92	58,115,014	28.09	100	Single end	SRR505744
		Flower	199,870,043	185,328,025	92.72	60,419,022	30.23	100	Single end	SRR656217
***P. patens* (Moss)**	Ppatens_251_v3	Whole plant	16,916,356	15,296,286	90.42	3,371,392	19.93	100	Paired-end	SRR787291
			23,179,054	22,023,877	95.02	4,377,357	18.88	100	Paired-end	SRR787292
			17,161,526	16,536,996	96.36	3,280,903	19.12	100	Paired-end	SRR787293
V. vinifera	Vvinifera_145	Young leaf	35,556,635	31,776,241	89.37	8,901,612	25.04	90	Single end	SRR2845691
		Mature leaf	38,432,356	33,860,317	88.10	9,517,790	24.77	90	Single end	SRR2845695
		Flower	46,674,468	40,010,518	85.72	11,701,136	25.07	100	Paired-end	SRR1302041

Table S4. RNA-seq coverage and alignment statistics of Brassicaceae

Species	Replicate	Tissue	Treatment	Sample Name	# trimmed reads	# uniquely mapped reads	% uniquely mapped reads
C. rubella	1	root	23°C	ruR1Crt	5,317,073	5,055,257	95.08
	2	root	23°C	ruR2Crt	7,904,012	7,533,558	95.31
	3	root	23°C	ruR3Crt	9,107,881	8,729,321	95.84
	1	shoot	23°C	ruR1Cst	7,333,968	6,985,621	95.25
	2	shoot	23°C	ruR2Cst	13,225,086	12,641,082	95.58
	3	shoot	23°C	ruR3Cst	17,800,088	16,913,304	95.02
	1	root	4°C	ruR1Trt	5,493,718	5,221,646	95.05
	2	root	4°C	ruR2Trt	6,416,316	6,110,719	95.24
	3	root	4°C	ruR3Trt	7,286,290	6,931,604	95.13
	1	shoot	4°C	ruR1Tst	13,416,367	12,794,728	95.37
	2	shoot	4°C	ruR2Tst	6,277,506	5,959,650	94.94
	3	shoot	4°C	ruR3Tst	6,544,658	6,219,569	95.03
A. lyrata	1	root	23°C	lyR1Crt	10,558,625	9,761,657	92.45
	2	root	23°C	lyR2Crt	6,190,542	5,703,305	92.13
	3	root	23°C	lyR3Crt	14,975,973	13,870,069	92.62
	1	shoot	23°C	lyR1Cst	6,918,348	6,352,992	91.83
	2	shoot	23°C	lyR2Cst	7,929,185	5,505,371	69.43
	3	shoot	23°C	lyR3Cst	5,698,224	5,171,457	90.76
	1	root	4°C	lyR1Trt	6,195,035	5,623,796	90.78
	2	root	4°C	lyR2Trt	5,429,237	4,987,578	91.87
	3	root	4°C	lyR3Trt	9,405,859	8,623,720	91.68
	1	shoot	4°C	lyR1Tst	10,896,846	9,988,297	91.66
	2	shoot	4°C	lyR2Tst	9,932,808	9,209,516	92.72
	3	shoot	4°C	lyR3Tst	8,638,965	7,990,118	92.49
A. thaliana	1	root	23°C	thR1Crt	10,818,845	10,355,200	95.71
	2	root	23°C	thR2Crt	12,676,650	12,135,294	95.73
	3	root	23°C	thR3Crt	13,671,238	13,152,185	96.20
	1	shoot	23°C	thR1Cst	13,623,199	12,922,671	94.86
	2	shoot	23°C	thR2Cst	8,303,406	7,944,093	95.67
	3	shoot	23°C	thR3Cst	16,318,352	15,591,912	95.55
	1	root	4°C	thR1Trt	8,356,115	7,870,059	94.18
	2	root	4°C	thR2Trt	5,865,111	3,934,492	67.08
	3	root	4°C	thR3Trt	7,244,682	7,007,661	96.73
	1	shoot	4°C	thR1Tst	10,249,337	9,818,160	95.79
	2	shoot	4°C	thR2Tst	16,609,325	15,878,152	95.60
	3	shoot	4°C	thR3Tst	13,203,435	12,638,423	95.72

4. General Discussion

Alternative splicing (AS) in plants has been studied for more than two decades and recent studies using next generation sequencing (NGS) technologies revealed that AS is prevalent in plants (Shen *et al.* 2014, Werneke *et al.* 1989, Wu *et al.* 2014). In several plant species, there are dynamic changes of AS under conditions of diverse abiotic stress, suggesting AS is closely linked with environmental stress responses in plants and may play an important role in plant adaptation and defense (Chang *et al.* 2014, Ding *et al.* 2014, Leviatan *et al.* 2013, Li *et al.* 2013, Marquez *et al.* 2012). However, the study of AS in response to biotic stress, especially insect herbivory attack remains lacking. Furthermore, two recent studies of the evolution of AS in animals proposed that the rapid divergence of AS may have played a more widespread role than divergence of mRNA expression in shaping species-specific differences (Barbosa-Morais *et al.* 2012, Merkin *et al.* 2012). In plants, however, such large scale comparative analyses on the evolution of AS, are missing, which prevent a deeper understanding of the functional role of AS in plants.

This dissertation describes the global changes of AS in plants in response to insect herbivory attack as well as the evolution of AS among different plant lineages. Specifically, in **manuscript I**, I investigated herbivory-induced genome-wide AS changes in both leaves and roots of *Nicotiana attenuata*, an ecological model plant for studying plant-herbivore interactions. I demonstrated that insect feeding can induce global AS responses in both leaves and roots, with stronger effect in roots than leaves. In addition, I found that induced AS responses in roots are likely due to jasmonic acid (JA) involved expression changes of SR protein coding genes. In **Manuscript II**, I annotated genome-wide alternative splicing in plants and found root, seed and anthers are the top three tissues that express the most tissue-specific splicing variants in *N. attenuata*. In **manuscript III**, I investigated the evolution of AS in plants using transcriptomic analysis of six plant species. I found that AS in plants evolves rapidly, and that divergence is largely due to gain and loss of AS events among orthologous genes. Furthermore, I revealed that the factors that determine alternative and constitutive splicing are conserved in plants, and the changes of these factors among different plant species largely contributes to the rapid turnover of AS in plants.

4.1 The function of environmental stress-induced AS in plants

Previous studies, including the work I presented in **manuscript I**, have revealed that both abiotic and biotic stresses can induce genome-wide AS changes in plants. However, whether these stress-induced AS changes are functional and may associate with acclimation response (functional hypothesis) or are simply a consequence of splicing errors caused by stresses (noise hypothesis) remains controversial (Chang, et al. 2014, Ding, et al. 2014, Ling *et al.* 2015, Reddy 2007, Thatcher *et al.* 2016). The noise hypothesis argues that most of the stress-induced AS events are just by-products of splicing errors. The rationale behind this hypothesis is: (i) although large amount of stress-induced AS events have been reported by several genome-wide studies, only a few cases have been shown to be functional (Mastrangelo *et al.* 2012); (ii) many stress-induced AS events may not be translated into functional proteins, since they introduce premature termination codon (PTC) into the resulting transcripts, and thus are targets of nonsense-mediated mRNA decay (NMD) (Ding, et al. 2014, Leviatan, et al. 2013); (iii) most stress-induced splicing variants are expressed at much lower levels in comparison to the dominant splicing isoforms (Cui and Xiong 2015); (iv) different stresses can activate the expression of a large number of genes which are related to plant stress-responses that are not expressed or have low expression under normal non-stressful conditions (Ding, et al. 2014, Ling, et al. 2015, Xiong *et al.* 2002, Yamaguchi-Shinozaki and Shinozaki 2006). Thus, there would be an urgent recruitment of a significant amount of splicing factors (SFs) to deal with the large amount of stress-induced pre-mRNAs. This would be a heavy burden for the whole splicing machinery, which may result in a large portion of these stress-induced pre-mRNAs failing to be processed properly; (v) in *A. thaliana*, the U6 snRNA, which is one of the core components of the spliceosome, has been found to be down-regulated under salt stress, which may affect the spliceosome assembly and its catalytic activity and cause splicing errors (Ding, et al. 2014).

On the other hand, the functional hypothesis that stress-induced AS is functional and plays an important role in contributing adaptation processes to stresses is also supported by different lines of evidence. Firstly, studies on individual genes revealed that the stress-induced AS in these genes are indeed functional (Liu *et al.* 2013, Matsukura *et al.* 2010, Qin *et al.* 2007, Yan *et al.* 2012). For example, in *A. thaliana*, *HsfA2*, the key regulator in response to heat stress, has three splicing variants (*HsfA2-I~III*). *HsfA2-III* is activated by severe heat and contains an in-

frame translational stop codon, and thus could result in a truncated protein with the length of only 129 amino acids (aa) when translated. Indeed, the truncated protein can be observed and was found to be able to bind to the promoter of *HsfA2* and activates its own transcription (Liu, et al. 2013). In addition, similar heat shock (HS) induced AS has also been detected in four other *Hsf* genes, indicating a common regulatory function of HS-induced AS exist in this gene family (Liu, et al. 2013). However, in comparison to the number of stress-induced AS identified by genome-wide analyses, there are only a handful of cases in which stress-induced AS has been functionally characterized. Therefore, the biological role of the many uncertified AS events induced by stress cannot be excluded. Secondly, that AS generates transcripts with PTC does not necessarily mean these are non-functional, as they can encode truncated proteins if they are not degraded. Indeed, many stress-induced transcripts are not degraded by NMD, despite having NMD features, suggesting some mechanisms exist to prevent these transcripts from being subjected to NMD (Leviatan, et al. 2013). Based on two recent studies, one of the possibilities is that these stress-induced intron-containing transcripts may have longer half-lives in the nucleus, thereby allow them to escape the NMD machinery and contribute to the regulation of gene expression (Boutz *et al.* 2015, Gohring *et al.* 2014). Furthermore, in several studies, the truncated proteins have been shown to have many functions, especially in response to stress (Dinesh-Kumar and Baker 2000, Liu, et al. 2013, Seo *et al.* 2011). In rice, *OsDREB2B*, a dehydration-responsive element binding gene underwent an exon skipping (ES) event at the second exon, generating two splicing variants either with or without PTC. Interestingly, the PTC+ transcript was more abundant than the PTC- transcript under non-stress conditions, while the relative abundance of the PTC- transcript increased under several stress conditions. The steady transcription of the PTC+ transcript under normal conditions suggests that it is a functional transcript (Matsukura, et al. 2010). Furthermore, the process of splicing variants being degraded by NMD can also be viewed as a functional process that controls the abundance of active protein at a post-transcriptional level. It has been demonstrated that the regions of AS that generate PTC in numerous SFs are ultra-conserved between different kingdoms, and that even if ancient unproductive splicing was lost in paralogs through gene duplications, its function would be rapidly and repeatedly replaced by newly raised distinct unproductive splicing (Lareau and Brenner 2015). This is further supported in **manuscript III**, which showed that PTC+ AS is more conserved than other AS in plants. Thirdly, some stress-induced AS of important

regulatory genes is conserved among diverse plant species (Filichkin *et al.* 2010). Finally, if stress-induced AS is a consequence of erroneous splicing caused by stress damage, we would expect AS changes to have no functional enrichment, since they likely occur randomly. However, several studies showed that a variety of stress induced AS events are involved in genes enriched in specific gene ontologies, such as RNA processing and stress responses (Chang, et al. 2014, Li, et al. 2013, Ling, et al. 2015). Taken together, this suggests that stress-induced AS deserves close investigation to determine its function before classifying them as artifacts or by-products of erroneous splicing. However, due to the large amount of splicing isoforms, it is difficult to characterize the function of each isoform. One possibility is to first narrow down the candidates using different strategies; for example, based on the conservation of the AS among diverse species or among populations of same species.

In **manuscript** I, I demonstrated that insect herbivory attack can elicit genome-wide transcriptomic alteration not only at the gene expression (GE) level but also at the pre-mRNA splicing level. More precisely, the regulation of GE and AS in response to herbivores is likely to be two parallel processes with independent functions. Based on several pieces of evidence, I believe induced AS is likely to be functional and involved in transcriptomic fine-tuning and anti-herbivore defenses or tolerance. First, unlike the situation described in previous studies, in which the majority of stress regulated AS introduces a PTC into the resulted transcripts (Ding, et al. 2014, Leviatan, et al. 2013), only a small fraction of herbivore-induced AS produces PTC+ transcripts and the proportion is even lower than the genome-wide average level, suggesting that these elicited AS may produce protein isoforms with distinct functions. Second, in both leaves and roots, specifically enriched gene ontologies (GOs) were observed in genes containing either herbivore-suppressed or elicited AS. Some of these enriched GO terms are known to be involved in direct or indirect anti-herbivore resistance in plants (Mitra and Baldwin 2008, Schwachtje and Baldwin 2008). Finally, the expressions of seven serine/arginine-rich (SR) genes, which are the key regulators of AS in plants, were up-regulated after herbivore attack in roots. It has been suggested by several studies that over-expression of certain SFs can increase plant tolerance to various stresses by significantly improving the splicing efficiency (Cui *et al.* 2014, Duque 2011, Forment *et al.* 2002). Furthermore, in *N. attenuata*, the root plays a vital role in many important processes including defense, resource reallocation and regrowth ability (Erb *et al.* 2009, Machado *et al.* 2013, Schwachtje *et al.* 2006, Steppuhn *et al.* 2004). All these results have

strengthened my hypothesis that the herbivore-induced AS is likely caused by precise transcriptomic regulation, which may function in anti-herbivore resistance or tolerance. However, further experiments are still required to validate this hypothesis.

4.2 Genome-wide AS alterations in response to environmental stresses in plants may be largely due to the regulation of SR proteins

In response to both biotic and abiotic stresses, rapid reprogramming of genome-wide pre-mRNA splicing was observed in different plant species. However, how environmental stresses change the global splicing pattern remains poorly understood. Studies have shown that members of the SR protein gene family are the key regulator of AS, which can affect splice site (SS) selection in a concentration- and phosphorylation-dependent manner, thus mediating AS regulation. Therefore, environmental stress-induced changes in SR genes may have large effects on the global alteration of AS (Duque 2011, Gao *et al.* 2004, Lopato *et al.* 1999a, Lopato *et al.* 1999b). In **manuscript I**, I demonstrated that insect feeding on leaves reduced the total abundance of AS in leaves (local tissue), but increased the total AS events in roots (systemic tissue). In addition, the number of identified differentially spliced (DS) gene in roots is also larger than the number in leaves. Interestingly, the expression of seven SR genes were significantly up-regulated in roots but not in leaves after herbivore attack, suggesting that the herbivore-induced global AS changes are largely associated with expression changes in SR genes. Indeed, several studies have revealed that overexpression of different SR genes in *A. thaliana* can influence pre-mRNA splicing of their own as well as some other genes (Kalyna *et al.* 2003, Simpson *et al.* 2008). The up-regulated expressions of specific SR genes were shown to be regulated by both abiotic stresses and abscisic acid (ABA), a major plant hormone with pivotal roles in response to various abiotic stresses (Cruz *et al.* 2014, Tanabe *et al.* 2007). Here, in response to herbivore attack, I found jasmonic acid (JA), another key phytohormone that mediates herbivory-induced transcriptome changes and defense responses (Kazan and Manners 2008, Kessler *et al.* 2004), is involved in the up-regulation of SR genes.

In addition to changes at the expression level, other regulation of SR genes might also be induced by stress, such as changes of splicing, medications of phosphorylation status and altered subcellular localizations, can all have a strong influence on the function of SR genes and thus contribute to genome-wide AS alterations. For example, the relative ratio of the transcript

encoding the full-length protein of *SR30*, which has been shown to affect the pre-mRNA splicing of its own, is dramatically increased under different stress conditions (Filichkin, et al. 2010, Lopato, et al. 1999b). Furthermore, one of the SR proteins in *A. thaliana* redistributed its preferential locations differentially in response to cold and heat stress in a phosphorylation status dependent manner (Ali *et al.* 2003).

In summary, genome-wide AS alterations in response to both biotic and abiotic stress in plants may largely depend on the stress-induced regulations of SR proteins. However, further experiments that directly manipulate the different status of SR proteins and investigate the AS pattern changes at individual gene level or genome-wide level are still required to draw more solid conclusions.

4.3 The splicing code in plants is largely conserved

The regulation of splice site choice is the key factor to determine whether a splicing junction will be constitutively spliced or alternatively spliced. The splicing code, a set of biological features that determine the splicing outcomes, has only recently been partially elucidated in animals but remains largely unknown in plants (Barash *et al.* 2010, Reddy *et al.* 2012). In **manuscript III**, using a machine learning approach, I analyzed the key determinants of AS in six different plant species. I demonstrated that the key determinants of AltA and AltD are highly conserved among plants. Although IR is the most abundant type of AS in plants (Kim *et al.* 2007, Marquez, et al. 2012), both AltA and AltD, which were attained by altering the position of splice sites (SS), also constitute a large portion of total AS among different eukaryote species (Kim, et al. 2007). For both AltA and AltD, studies in mammals showed that the distance between the alternative SS (minor form) and authentic SS (major form) is highly diverse (Dou *et al.* 2006, Koren *et al.* 2007), but the frequency of alternative SS usage is the greatest close to the authentic SS (within 6 nt), and drops farther away from the authentic SS (Dou, et al. 2006). This pattern is consistent with two recent studies, which demonstrated that the closer alternative SS are more likely to be used (Barash, et al. 2010, Rosenberg *et al.* 2015). In **manuscript III**, I also found that the distance between the nearest 5'/3' alternative SS and the authentic SS is one of the most important factors in determining AltD/AltA, suggesting that the regulation of AS is at least partially conserved between plants and mammals. Furthermore, whether the putative alternative SS would preserve the original protein reading frame is also an important factor that contributed

to the determination of AltD and AltA. This is consistent with previous analysis that demonstrated a prominent in-frame bias for alternative 5' and 3' SS usage in both plants and mammals (Dou, et al. 2006, Hiller *et al.* 2004, Satyawan *et al.* 2016).

By comparing constitutively spliced exons with alternative 5' and 3' exons (A5Es and A3Es) in humans, several features that are distinct between the two exon groups have been identified (Koren, et al. 2007). First, both A5Es and A3Es demonstrate a high strength SS in the constitutively spliced side (3' SS for A5Es and 5'SS for A3Es) and a weak SS in the alternative spliced side, thus providing a strong anchor at the constitutive SS and suboptimal SS at the other side. Second, for AltA, the major SS was always stronger than the minor SS upstream or downstream. However, for AltD, if the minor SS is located downstream of the major SS, there is no significant difference between the two SS in terms of strength (Koren, et al. 2007). Interestingly, they also found in such situations, that the upstream region of the major SS has a higher exon splicing enhancer (ESE) density than that of the minor SS and exon splicing silencer (ESS) showed an opposite pattern. However, no significant difference was found between major and minor SS for AltA in terms of exon splicing regulator (ESR) (Koren, et al. 2007). These results suggest that ESR may play a more important role in the proper selection of alternative SS in AltD than AltA, and the alternative 3' SS is mainly due to the downstream screening of the branch site (BS) and polypyrimidine tract (PPT) (Smith *et al.* 1993, Smith *et al.* 1989). This may explain why in **manuscript III**, our AltD model had an overall lower precision rate than the model of AltA. Although we attempted to identify a few sequence motifs that may act as putative splicing regulators, none of them contributed to the model prediction, suggesting the identified sequence motifs might not be the splicing regulators that are regulating AltD. Future research that applies different approaches, such as introducing millions of random hexamers into mini-genes and investigate the consequences of AS, will provide more insight on splicing regulators involved in AltD regulations in plants.

IR is the most common AS type in plants and the least prevalent in animals (Kim, et al. 2007). Although, IR is known to play a role in many important regulatory processes, such as mRNA export and cellular localizations, which often coupled with NMD (Filichkin, et al. 2010, Li *et al.* 2006), the mechanisms that regulate of IR are the least understood in comparison to other types of AS. More than a decade ago, a hypothesis proposed that for IR, the competition is between splicing and mRNA transport rather than between two SS. Because transcripts with IR,

as partially spliced RNA products must be exported to the cytoplasm, how the cell judges if the pre-mRNA splicing is complete and ready for moving to the cytoplasm is important (Black 2003). Indeed, it was shown in a recent study that many introns, which are recognized as 'retained introns', have long half-lives and remain in the nucleus, thus avoiding be subjected to NMD (Boutz, et al. 2015). Furthermore, the retained introns are shown to be significantly associated with increased GC content, reduced intron length, relatively weak SS and a reduced availability of the spliceosome caused by localized stalling of RNA polymerase II (Braunschweig *et al.* 2014). In **manuscript III**, I have also demonstrated that junctions with IR are significantly shorter and have weaker 5' and 3' SS than constitutive spliced junctions in plants. However, these features alone cannot result in the accurate prediction of IR, suggesting that other key determinants were not detected in our study. In addition, it is possible that the sequence depth of the data we used was not sufficient to separate IR from constitutive splicing.

Exon skipping (ES) is the least prevalent AS type in plants but the most common in animals (Kim, et al. 2007). The regulation of ES in animal models is also closely related to the SS strength of both 5' and 3', the BS and *cis*-regulatory elements (Koren, et al. 2007, Lev-Maor *et al.* 2007, Miriami *et al.* 2003). The regulation of ES in plants remains largely unknown due to its rare occurrence.

4.4 The rapid evolution of AS in plants is mainly *cis*-regulated

In **manuscript III**, I demonstrated that the profiles of AS evolved rapidly in plants, which may be largely due to gain and loss of AS events among orthologous genes. This observation is similar to the pattern reported in two vertebrates studies spanning ~350 million years, which also showed a rapid evolution of AS (Barbosa-Morais, et al. 2012, Merkin, et al. 2012). Specifically, they demonstrated that AS is largely species-specific while the overall GE pattern is segregated according to tissue types, suggesting species-specific AS has evolved much faster than species-specific GE, and therefore lineage-specific AS may be largely associated with speciation (Barbosa-Morais, et al. 2012, Merkin, et al. 2012). In **manuscript III**, we also found a dominant species-specific pattern of AS and this pattern is consistent even among closely related species, suggesting the rapid evolution of AS in plants. Interestingly, in comparison to vertebrates, a distinct pattern of GE was observed in plants, of which tissue- dominant clustering was only observed among closely related species. This observation is consistent with a previous

study that focused on three highly divergent plants including both monocots and dicots (diverged ~200 million years ago). The authors demonstrated that the global pattern of GE is grouped according to species rather than tissues (Yang and Wang 2013). Together, both our data and previous studies indicate that ranscriptome evolution rates in plants, at both the GE and the AS level, are faster than in vertebrates. Furthermore, aside from the dominant species-specific AS, some tissues in vertebrates, such as the brain, testis, heart and muscle still show strong tissue-splicing signatures (Barbosa-Morais, et al. 2012, Merkin, et al. 2012). Based on the three major tissues (root, leaves and flowers) we analyzed, we did not find such pattern. However, studies showed that the transcriptomes of sexual tissues are substantially different from those of vegetative tissues, and the anthers harbor the most diverged specialized metabolome (Li *et al.* 2016, Yang and Wang 2013). Therefore, future research including transcriptome data of much diverse tissues may provide a more conclusive answer.

Species-specific AS profiles may have contributed to the phenotypic diversity among species. It has been demonstrated that a subset of splicing that is highly diverged among different vertebrate species, was predicted to remodel protein to protein interactions (PPIs) and act as trans-acting regulators (Barbosa-Morais, et al. 2012). This may further contribute to the diversification of other transcriptomic changes including splicing and the underlying phenotypic traits (Barbosa-Morais, et al. 2012). In plants, to what extent rapid AS changes contribute to the phenotypic divergence among different species remains an open question. Future studies focused on identifying species-specific AS that are also under positive selection within the populations will shed light on this aspect.

In **manuscript III**, I not only identified several key determinants of AltD/AltA that are highly conserved in plants but also showed that the variations of these AS determinants largely contributed to the AS evolution between closely related species. Interestingly, most of the key determinants such as SS, and *cis*-elements (BS, PPT, UA-rich tract) are all located in the intronic region. It is known that sequences of intron evolve rapidly (Hare and Palumbi 2003, Mattick 1994), a process which is likely to have resulted in rapid gain and loss of AS between different lineages. Our results are also consistent with findings in mammals, in which the mutations affecting intronic splicing regulatory elements (ISREs) were shown to be the main factor that caused distinct splicing patterns between different lineages (Merkin, et al. 2012). Our results also support the model that alternatively spliced exons originated from constitutive spliced exons

though weakening of the SS or the introducing of a competing SS nearby (Lev-Maor, et al. 2007), as changes in the distance between authentic SS and alternative SS is negatively associated with the conservation of AS.

During evolution, if any mutations/insertions took place near SS or SRE regions, the effects on the splicing will be observed in all tissues. This explains, at least partially, the observed species-specific AS profiles in both plants and animals. However, we cannot exclude the possibility that the species-specific trans-factors, such as SR protein family, which have different numbers of homologues among species (Iida and Go 2006, Isshiki *et al.* 2006, Ling, et al. 2015), may have also contributed to the divergence of AS among different species (Ast 2004, Barbosa-Morais, et al. 2012).

Several other factors such as gene duplication (GD) DNA methylations and transposable element (TE) insertions have also been hypothesized to affect AS evolution (Flores *et al.* 2012, Sorek *et al.* 2002, Su *et al.* 2006). In **manuscript III** I investigated all of these factors and found that most of them (except TE insertions, the effects of which were found to be specific-specific) did not significantly contribute to AS conservation between species. The species-specific effects of TE on AS conservation are largely due to the different abundance of TE insertions in the genomes of different species (Hu *et al.* 2011, Sierro *et al.* 2014, Slotte *et al.* 2013, Tomato Genome 2012), suggesting genomic composition of each species might also affect the evolutionary trajectory of AS.

Taken together, using comparative and genome-wide analysis, my dissertation provides both evolutionary and mechanistic insights on how AS were regulated by environmental stresses and the underlying codes that contributed to the rapid divergence of AS among plants. It is clear that much more functional characterizations on AS regulation machinery as well as specific alternatively spliced transcripts in plants are urgently needed, and for which the AS determinants I identified will provide both valuable indications and a theoretical framework.

References

Ali, G.S., Golovkin, M. and Reddy, A.S. (2003) Nuclear localization and in vivo dynamics of a plant-specific serine/arginine-rich protein. *Plant J*, **36**, 883-893.

Ast, G. (2004) How did alternative splicing evolve? *Nat Rev Genet*, **5**, 773-782.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., et al. (2010) Deciphering the splicing code. *Nature*, **465**, 53-59.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., et al. (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587-1593.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**, 291-336.

Boutz, P.L., Bhutkar, A. and Sharp, P.A. (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev*, **29**, 63-80.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., et al. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res*, **24**, 1774-1786.

Chang, C.Y., Lin, W.D. and Tu, S.L. (2014) Genome-Wide Analysis of Heat-Sensitive Alternative Splicing in *Physcomitrella patens*. *Plant Physiol*, **165**, 826-840.

Cruz, T.M., Carvalho, R.F., Richardson, D.N. and Duque, P. (2014) Abscisic acid (ABA) regulation of Arabidopsis SR protein gene expression. *Int J Mol Sci*, **15**, 17541-17564.

Cui, P. and Xiong, L. (2015) Environmental Stress and Pre-mRNA Splicing. *Mol Plant*, **8**, 1302-1303.

Cui, P., Zhang, S., Ding, F., Ali, S. and Xiong, L. (2014) Dynamic regulation of genome-wide pre-mRNA splicing and stress tolerance by the Sm-like protein LSm5 in Arabidopsis. *Genome Biol*, **15**, R1.

Dinesh-Kumar, S.P. and Baker, B.J. (2000) Alternatively spliced N resistance gene transcripts: their possible role in tobacco mosaic virus resistance. *Proc Natl Acad Sci U S A*, **97**, 1908-1913.

Ding, F., Cui, P., Wang, Z., Zhang, S., Ali, S. and Xiong, L. (2014) Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. *BMC Genomics*, **15**, 431.

Dou, Y., Fox-Walsh, K.L., Baldi, P.F. and Hertel, K.J. (2006) Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*, **12**, 2047-2056.

Duque, P. (2011) A role for SR proteins in plant stress responses. *Plant Signal Behav*, **6**, 49-54.

Erb, M., Lenk, C., Degenhardt, J. and Turlings, T.C. (2009) The underestimated role of roots in defense against leaf attackers. *Trends Plant Sci*, **14**, 653-659.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., et al. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res*, **20**, 45-58.

Flores, K., Wolschin, F., Corneveaux, J.J., Allen, A.N., Huentelman, M.J. and Amdam, G.V. (2012) Genome-wide association between DNA methylation and alternative splicing in an invertebrate. *BMC Genomics*, **13**, 480.

Forment, J., Naranjo, M.A., Roldan, M., Serrano, R. and Vicente, O. (2002) Expression of Arabidopsis SR-like splicing proteins confers salt tolerance to yeast and transgenic plants. *Plant J*, **30**, 511-519.

Gao, H., Gordon-Kamm, W.J. and Lyznik, L.A. (2004) ASF/SF2-like maize pre-mRNA splicing factors affect splice site utilization and their transcripts are alternatively spliced. *Gene*, **339**, 25-37.

Gohring, J., Jacak, J. and Barta, A. (2014) Imaging of endogenous messenger RNA splice variants in living cells reveals nuclear retention of transcripts inaccessible to nonsense-mediated decay in Arabidopsis. *Plant Cell*, **26**, 754-764.

Hare, M.P. and Palumbi, S.R. (2003) High intron sequence conservation across three mammalian orders suggests functional constraints. *Mol Biol Evol*, **20**, 969-978.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., et al. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet*, **36**, 1255-1257.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., et al. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet*, **43**, 476-481.

Iida, K. and Go, M. (2006) Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. *Mol Biol Evol*, **23**, 1085-1094.

Isshiki, M., Tsumoto, A. and Shimamoto, K. (2006) The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. *Plant Cell*, **18**, 146-158.

Kalyna, M., Lopato, S. and Barta, A. (2003) Ectopic expression of atRSZ33 reveals its function in splicing and causes pleiotropic changes in development. *Mol Biol Cell*, **14**, 3565-3577.

Kazan, K. and Manners, J.M. (2008) Jasmonate signaling: toward an integrated view. *Plant Physiol*, **146**, 1459-1468.

Kessler, A., Halitschke, R. and Baldwin, I.T. (2004) Silencing the jasmonate cascade: induced plant defenses and insect populations. *Science*, **305**, 665-668.

Kim, E., Magen, A. and Ast, G. (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, **35**, 125-131.

Koren, E., Lev-Maor, G. and Ast, G. (2007) The emergence of alternative 3' and 5' splice site exons from constitutive exons. *PLoS Comput Biol*, **3**, e95.

Lareau, L.F. and Brenner, S.E. (2015) Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol Biol Evol*, **32**, 1072-1079.

Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., et al. (2007) The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet*, **3**, e203.

Leviatan, N., Alkan, N., Leshkowitz, D. and Fluhr, R. (2013) Genome-wide survey of cold stress regulated alternative splicing in *Arabidopsis thaliana* with tiling microarray. *PLoS One*, **8**, e66511.

Li, D., Heiling, S., Baldwin, I.T. and Gaquerel, E. (2016) Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. *Proc Natl Acad Sci U S A*.

Li, W., Lin, W.D., Ray, P., Lan, P. and Schmidt, W. (2013) Genome-wide detection of condition-sensitive alternative splicing in Arabidopsis roots. *Plant Physiol*, **162**, 1750-1763.

Li, Y., Bor, Y.C., Misawa, Y., Xue, Y., Rekosh, D. and Hammarskjold, M.L. (2006) An intron with a constitutive transport element is retained in a Tap messenger RNA. *Nature*, **443**, 234-237.

Ling, Z., Zhou, W., Baldwin, I.T. and Xu, S. (2015) Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*. *Plant J*, **84**, 228-243.

Liu, J., Sun, N., Liu, M., Liu, J., Du, B., Wang, X., et al. (2013) An autoregulatory loop controlling Arabidopsis HsfA2 expression: role of heat shock-induced alternative splicing. *Plant Physiol*, **162**, 512-521.

Lopato, S., Gattoni, R., Fabini, G., Stevenin, J. and Barta, A. (1999a) A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. *Plant Mol Biol*, **39**, 761-773.

Lopato, S., Kalyna, M., Dorner, S., Kobayashi, R., Krainer, A.R. and Barta, A. (1999b) atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes. *Genes Dev*, **13**, 987-1001.

Machado, R.A., Ferrieri, A.P., Robert, C.A., Glauser, G., Kallenbach, M., Baldwin, I.T., et al. (2013) Leaf-herbivore attack reduces carbon reserves and regrowth from the roots via jasmonate and auxin signaling. *New Phytol*, **200**, 1234-1246.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res*, **22**, 1184-1195.

Mastrangelo, A.M., Marone, D., Laido, G., De Leonardis, A.M. and De Vita, P. (2012) Alternative splicing: enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci*, **185-186**, 40-49.

Matsukura, S., Mizoi, J., Yoshida, T., Todaka, D., Ito, Y., Maruyama, K., et al. (2010) Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes. *Mol Genet Genomics*, **283**, 185-196.

Mattick, J.S. (1994) Introns: evolution and function. *Curr Opin Genet Dev*, **4**, 823-831.

Merkin, J., Russell, C., Chen, P. and Burge, C.B. (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593-1599.

Miriami, E., Margalit, H. and Sperling, R. (2003) Conserved sequence elements associated with exon skipping. *Nucleic Acids Res*, **31**, 1974-1983.

Mitra, S. and Baldwin, I.T. (2008) Independently silencing two photosynthetic proteins in *Nicotiana attenuata* has different effects on herbivore resistance. *Plant Physiol*, **148**, 1128-1138.

Qin, F., Kakimoto, M., Sakuma, Y., Maruyama, K., Osakabe, Y., Tran, L.S., et al. (2007) Regulation and functional analysis of ZmDREB2A in response to drought and heat stresses in *Zea mays* L. *Plant J*, **50**, 54-69.

Reddy, A.S. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol*, **58**, 267-294.

Reddy, A.S., Rogers, M.F., Richardson, D.N., Hamilton, M. and Ben-Hur, A. (2012) Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Front Plant Sci*, **3**, 18.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J. and Seelig, G. (2015) Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, **163**, 698-711.

Satyawan, D., Kim, M.Y. and Lee, S.H. (2016) Stochastic alternative splicing is prevalent in mungbean (*Vigna radiata*). *Plant Biotechnol J*.

Schwachtje, J. and Baldwin, I.T. (2008) Why does herbivore attack reconfigure primary metabolism? *Plant Physiol*, **146**, 845-851.

Schwachtje, J., Minchin, P.E., Jahnke, S., van Dongen, J.T., Schittko, U. and Baldwin, I.T. (2006) SNF1-related kinases allow plants to tolerate herbivory by allocating carbon to roots. *Proc Natl Acad Sci U S A*, **103**, 12935-12940.

Seo, P.J., Kim, M.J., Ryu, J.Y., Jeong, E.Y. and Park, C.M. (2011) Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. *Nat Commun*, **2**, 303.

Shen, Y., Zhou, Z., Wang, Z., Li, W., Fang, C., Wu, M., et al. (2014) Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell*, **26**, 996-1008.

Sierro, N., Battey, J.N., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., et al. (2014) The tobacco genome sequence and its comparison with those of tomato and potato. *Nat Commun*, **5**, 3833.

Simpson, C.G., Fuller, J., Maronova, M., Kalyna, M., Davidson, D., McNicol, J., et al. (2008) Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant J*, **53**, 1035-1048.

Slotte, T., Hazzouri, K.M., Agren, J.A., Koenig, D., Maumus, F., Guo, Y.L., et al. (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet*, **45**, 831-835.

Smith, C.W., Chu, T.T. and Nadal-Ginard, B. (1993) Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol*, **13**, 4939-4952.

Smith, C.W., Porro, E.B., Patton, J.G. and Nadal-Ginard, B. (1989) Scanning from an independently specified branch point defines the 3' splice site of mammalian introns. *Nature*, **342**, 243-247.

Sorek, R., Ast, G. and Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res*, **12**, 1060-1067.

Steppuhn, A., Gase, K., Krock, B., Halitschke, R. and Baldwin, I.T. (2004) Nicotine's defensive function in nature. *PLoS Biol*, **2**, E217.

Su, Z., Wang, J., Yu, J., Huang, X. and Gu, X. (2006) Evolution of alternative splicing after gene duplication. *Genome Res*, **16**, 182-189.

Tanabe, N., Yoshimura, K., Kimura, A., Yabuta, Y. and Shigeoka, S. (2007) Differential expression of alternatively spliced mRNAs of Arabidopsis SR protein homologs, atSR30 and atSR45a, in response to environmental stress. *Plant Cell Physiol*, **48**, 1036-1049.

Thatcher, S.R., Danilevskaya, O.N., Meng, X., Beatty, M., Zastrow-Hayes, G., Harris, C., et al. (2016) Genome-Wide Analysis of Alternative Splicing during Development and Drought Stress in Maize. *Plant Physiol*, **170**, 586-599.

Tomato Genome, C. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635-641.

Werneke, J.M., Chatfield, J.M. and Ogren, W.L. (1989) Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and Arabidopsis. *Plant Cell*, **1**, 815-825.

Wu, H.P., Su, Y.S., Chen, H.C., Chen, Y.R., Wu, C.C., Lin, W.D., et al. (2014) Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in *Physcomitrella patens*. *Genome Biol*, **15**, R10.

Xiong, L., Schumaker, K.S. and Zhu, J.K. (2002) Cell signaling during cold, drought, and salt stress. *Plant Cell*, **14 Suppl**, S165-183.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (2006) Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu Rev Plant Biol*, **57**, 781-803.

Yan, K., Liu, P., Wu, C.A., Yang, G.D., Xu, R., Guo, Q.H., et al. (2012) Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in *Arabidopsis thaliana*. *Mol Cell*, **48**, 521-531.

Yang, R. and Wang, X. (2013) Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *Plant Cell*, **25**, 71-82.

Summary

Recent studies showed that alternative splicing (AS) is much more prevalent in plants than previously thought. Both genome-wide analyses and cases studies of AS in response to abiotic stresses indicate the importance of AS for plant adaptation in a changing environment. However, how plants regulate their AS in response to herbivore attack, one of major biotic stresses that threaten plant fitness, remains unknown. The tight association between AS and environmental stresses also points to the rapid evolution of AS. However, the underlying mechanisms for this rapid evolution are unknown in plants.

In this thesis, I aimed to address these two main questions by investigating the genome-wide insect herbivore-induced AS alteration in wild tobacco (*Nicotiana attenuata*) and by systematically studying the mechanisms that contributed to the evolution of AS among six plant species. For the first question, I found that insect herbivory elicits genome-wide pre-mRNA splicing changes in local tissue (leaves), but also in systemic tissue (roots) of *N. attenuata*. Interestingly, the herbivory-induced AS alterations are stronger in systemic tissues than in local tissues. This process is likely due to jasmonic acid (JA) dependent upregulation of several serine/arginine-rich (SR) protein encoded genes. Functional annotations of AS genes suggest that herbivory-induced AS responses of both leaves and roots are likely to be involved in anti-herbivore defenses and resistance. For the second question, I found AS evolved rapidly among different plant species, as the AS profiles are more similar among different tissues within the species than among the same tissue between different species. This process is largely due to the rapid gain and loss of AS among orthologous genes. Using a machine learning approach, I found that the determinants of AS are largely conserved among studied plant species and variations of these determinants largely contributed to the rapid changes of AS among species. Despite rapid turnover of AS among species, I also found several ultra-conserved AS events in plants, most of which likely resulted in introducing premature termination codon (PTC) containing transcripts and are likely subject to nonsense-mediated decay (NMD), suggesting that AS coupled with nonsense-mediated decay (NMD) is conserved and plays an important role at the post-transcriptional level in plants.

In conclusion, my thesis provides evidence that insect herbivores can induce genome-wide responses of AS in plants and those changes of key AS determinants contributed to the

rapid evolution of AS in plants. In addition, this thesis provides abundant resources for studying AS functions and a computational framework for investigating AS evolution in plants.

Zusammenfassung

Neueste Forschungsergebnisse zeigen, dass alternatives Spleißen in Pflanzen weitaus verbreiteter ist, als ursprünglich angenommen. Fallbeispiele sowie genomweite Untersuchungen weisen auf die Bedeutung des alternativen Spleißens für die Anpassung der Pflanzen als Antwort auf abiotischen Stress in einer sich verändernden Umwelt hin. Allerdings ist bis heute unbekannt, wie Pflanzen ihr alternatives Spleißen während des Befalls durch Pflanzenfresser regulieren. Des Weiteren deutet die enge Verbindung zwischen alternativem Spleißen und biotischen sowie abiotischen Umwelteinflüssen darauf hin, dass sich alternatives Spleißen evolutionär schnell entwickelt. Trotzdem sind die zugrunde liegenden Mechanismen in Pflanzen weitgehend unbekannt.

Das Ziel der vorliegenden Dissertation ist es, auf diese beiden Punkte einzugehen. Hierfür wurden die genomweiten Veränderungen des durch pflanzenfressende Insekten ausgelösten alternativen Spleißens in *Nicotiana attenuata* untersucht. Des Weiteren wurden die Mechanismen in sechs weiteren Arten analysiert, die zur Evolution des alternativen Spleißens beitragen. In *N. attenuata* lösen pflanzenfressende Insekten in lokal befallenem (Blätter) als auch in systemisch betroffenem Gewebe (Wurzeln) eine genomweite Veränderung des Spleißens der prä-mRNA aus. Interessanterweise fällt hierbei die Veränderung in systemischem Gewebe stärker aus als in lokalem Gewebe. Die Ursache hierfür könnte die erhöhte Jasmonsäure-abhängige Expression Serin/Arginin-reicher proteinkodierender Gene sein. Die funktionellen Vorhersagen der Gene für alternatives Spleißen weisen darauf hin, dass die durch Pflanzenfresser ausgelöste Reaktion des alternativen Spleißens in Blättern und Wurzeln an deren Abwehr und Toleranz beteiligt ist. Alternatives Spleißen entwickelt sich in evolutionären Zeiträumen rasant in verschiedenen Pflanzenarten. Dabei sind die Profile verschiedener Gewebe einer Art ähnlicher als die Profile von gleichen Geweben verschiedener Arten. Der Grund dafür liegt wahrscheinlich in der schnellen Zu- und Abnahme des alternativen Spleißens in orthologen Genen. Unter Verwendung von maschinellem Lernen konnte gezeigt werden, dass die bestimmenden Faktoren für alternatives Spleißen in den ausgewählten Pflanzenarten weitgehend erhalten sind. Die Streuung dieser Faktoren in den unterschiedlichen Arten trägt dabei maßgeblich zu den schnellen Veränderungen des alternativen Spleißens bei. Ungeachtet dieser

Fluktuationen zwischen den Arten konnten stark konservierte alternative Spleißereignisse festgestellt werden. Wahrscheinlich führten die meisten dieser Ereignisse zur Einführung eines vorzeitigen Stopp Codons im offenen Leseraster einer mRNA und lösten damit den Nonsense-mediated mRNA Decay (NMD) aus. Das weist wiederum darauf hin, dass die Kopplung zwischen alternativem Spleißen und NMD konserviert ist und eine bedeutende Rolle auf post-transkriptionaler Ebene in Pflanzen spielt.

Zusammenfassend kann festgestellt werden, dass die vorliegende Dissertation Beweise liefert, dass pflanzenfressende Insekten eine genomweite Reaktion des alternativen Spleißens auslösen können. Des Weiteren steuern Veränderungen bestimmter Schlüsselfaktoren für alternatives Spleißen zur raschen Entwicklung dieser bei. Zusätzlich liefert diese Dissertation ergiebige Bezugsquellen, um die Funktion des alternativen Spleißens zu untersuchen, sowie eine computergestützte Analyseplattform für die Erforschung der Evolution des alternativen Spleißens in Pflanzen.

Bibliography

Ali, G.S., Golovkin, M. and Reddy, A.S. (2003) Nuclear localization and in vivo dynamics of a plant-specific serine/arginine-rich protein. *Plant J*, **36**, 883-893.

Ali, G.S. and Reddy, A.S. (2008a) Regulation of alternative splicing of pre-mRNAs by stresses. *Curr Top Microbiol Immunol*, **326**, 257-275.

Ali, G.S. and Reddy, A.S. (2008b) Spatiotemporal organization of pre-mRNA splicing proteins in plants. *Curr Top Microbiol Immunol*, **326**, 103-118.

Arciga-Reyes, L., Wootton, L., Kieffer, M. and Davies, B. (2006) UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in Arabidopsis. *Plant J*, **47**, 480-489.

Ast, G. (2004) How did alternative splicing evolve? *Nat Rev Genet*, **5**, 773-782.

Baek, J.M., Han, P., Iandolino, A. and Cook, D.R. (2008) Characterization and comparison of intron structure and alternative splicing between *Medicago truncatula*, *Populus trichocarpa*, Arabidopsis and rice. *Plant Mol Biol*, **67**, 499-510.

Bahulikar, R.A., Stanculescu, D., Preston, C.A. and Baldwin, I.T. (2004) ISSR and AFLP analysis of the temporal and spatial population structure of the post-fire annual, *Nicotiana attenuata*, in SW Utah. *BMC Ecol*, **4**, 12.

Baldwin, I.T., Staszak-Kozinski, L. and Davidson, R. (1994) Up in smoke: I. Smoke-derived germination cues for postfire annual,*Nicotiana attenuata* torr. Ex. Watson. *J Chem Ecol*, **20**, 2345-2371.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., et al. (2010) Deciphering the splicing code. *Nature*, **465**, 53-59.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., et al. (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587-1593.

Barrass, J.D. and Beggs, J.D. (2003) Splicing goes global. *Trends Genet*, **19**, 295-298.

Barta, A., Sommergruber, K., Thompson, D., Hartmuth, K., Matzke, M.A. and Matzke, A.J. (1986) The expression of a nopaline synthase - human growth hormone chimaeric gene in transformed tobacco and sunflower callus tissue. *Plant Mol Biol*, **6**, 347-357.

Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215-233.

Belostotsky, D.A. and Sieburth, L.E. (2009) Kill the messenger: mRNA decay and plant development. *Curr Opin Plant Biol*, **12**, 96-102.

Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J Biol Chem*, **270**, 2411-2414.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**, 291-336.

Boutz, P.L., Bhutkar, A. and Sharp, P.A. (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev*, **29**, 63-80.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., et al. (2014) Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res*, **24**, 1774-1786.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J. and Bork, P. (2002) Alternative splicing and genome complexity. *Nat Genet*, **30**, 29-30.

Brown, J.W. (1996) Arabidopsis intron mutations and pre-mRNA splicing. *Plant J*, **10**, 771-780.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. and Buell, C.R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics*, **7**, 327.

Capovilla, G., Pajoro, A., Immink, R.G. and Schmid, M. (2015) Role of alternative pre-mRNA splicing in temperature signaling. *Curr Opin Plant Biol*, **27**, 97-103.

Chamala, S., Feng, G., Chavarro, C. and Barbazuk, W.B. (2015) Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. *Front Bioeng Biotechnol*, **3**, 33.

Chang, C.Y., Lin, W.D. and Tu, S.L. (2014) Genome-Wide Analysis of Heat-Sensitive Alternative Splicing in *Physcomitrella patens*. *Plant Physiol*, **165**, 826-840.

Chang, Y.F., Imam, J.S. and Wilkinson, M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*, **76**, 51-74.

Chen, M. and Manley, J.L. (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*, **10**, 741-754.

Cruz, T.M., Carvalho, R.F., Richardson, D.N. and Duque, P. (2014) Abscisic acid (ABA) regulation of Arabidopsis SR protein gene expression. *Int J Mol Sci*, **15**, 17541-17564.

Cui, P. and Xiong, L. (2015) Environmental Stress and Pre-mRNA Splicing. *Mol Plant*, **8**, 1302-1303.

Cui, P., Zhang, S., Ding, F., Ali, S. and Xiong, L. (2014) Dynamic regulation of genome-wide pre-mRNA splicing and stress tolerance by the Sm-like protein LSm5 in Arabidopsis. *Genome Biol*, **15**, R1.

Darracq, A. and Adams, K.L. (2013) Features of evolutionarily conserved alternative splicing events between Brassica and Arabidopsis. *New Phytol*, **199**, 252-263.

de la Fuente van Bentem, S., Anrather, D., Dohnal, I., Roitinger, E., Csaszar, E., Joore, J., et al. (2008) Site-specific phosphorylation profiling of Arabidopsis proteins by mass spectrometry and peptide chip analysis. *J Proteome Res*, **7**, 2458-2470.

Dinesh-Kumar, S.P. and Baker, B.J. (2000) Alternatively spliced N resistance gene transcripts: their possible role in tobacco mosaic virus resistance. *Proc Natl Acad Sci U S A*, **97**, 1908-1913.

Ding, F., Cui, P., Wang, Z., Zhang, S., Ali, S. and Xiong, L. (2014) Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. *BMC Genomics*, **15**, 431.

Donahue, C.P., Muratore, C., Wu, J.Y., Kosik, K.S. and Wolfe, M.S. (2006) Stabilization of the tau exon 10 stem loop alters pre-mRNA splicing. *J Biol Chem*, **281**, 23302-23306.

Dou, Y., Fox-Walsh, K.L., Baldi, P.F. and Hertel, K.J. (2006) Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*, **12**, 2047-2056.

Duque, P. (2011) A role for SR proteins in plant stress responses. *Plant Signal Behav*, **6**, 49-54.

Eckardt, N.A. (2002) Alternative splicing and the control of flowering time. *Plant Cell*, **14**, 743-747.

Erb, M., Lenk, C., Degenhardt, J. and Turlings, T.C. (2009) The underestimated role of roots in defense against leaf attackers. *Trends Plant Sci*, **14**, 653-659.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., et al. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res*, **20**, 45-58.

Flores, K., Wolschin, F., Corneveaux, J.J., Allen, A.N., Huentelman, M.J. and Amdam, G.V. (2012) Genome-wide association between DNA methylation and alternative splicing in an invertebrate. *BMC Genomics*, **13**, 480.

Forment, J., Naranjo, M.A., Roldan, M., Serrano, R. and Vicente, O. (2002) Expression of Arabidopsis SR-like splicing proteins confers salt tolerance to yeast and transgenic plants. *Plant J*, **30**, 511-519.

Gao, H., Gordon-Kamm, W.J. and Lyznik, L.A. (2004) ASF/SF2-like maize pre-mRNA splicing factors affect splice site utilization and their transcripts are alternatively spliced. *Gene*, **339**, 25-37.

Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.

Gohring, J., Jacak, J. and Barta, A. (2014) Imaging of endogenous messenger RNA splice variants in living cells reveals nuclear retention of transcripts inaccessible to nonsense-mediated decay in Arabidopsis. *Plant Cell*, **26**, 754-764.

Gulledge, A.A., Roberts, A.D., Vora, H., Patel, K. and Loraine, A.E. (2012) Mining *Arabidopsis thaliana* RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. *Am J Bot*, **99**, 219-231.

Hare, M.P. and Palumbi, S.R. (2003) High intron sequence conservation across three mammalian orders suggests functional constraints. *Mol Biol Evol*, **20**, 969-978.

Hartmuth, K. and Barta, A. (1986) In vitro processing of a plant pre-mRNA in a HeLa cell nuclear extract. *Nucleic Acids Res*, **14**, 7513-7528.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., et al. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet*, **36**, 1255-1257.

Hori, K. and Watanabe, Y. (2005) UPF3 suppresses aberrant spliced mRNA in Arabidopsis. *Plant J*, **43**, 530-540.

Howard, B.E., Hu, Q., Babaoglu, A.C., Chandra, M., Borghi, M., Tan, X., et al. (2013) High-throughput RNA sequencing of pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants. *PLoS One*, **8**, e74183.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., et al. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet*, **43**, 476-481.

Iida, K. and Go, M. (2006) Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. *Mol Biol Evol*, **23**, 1085-1094.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., et al. (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res*, **32**, 5096-5103.

Iida, K., Shionyu, M. and Suso, Y. (2008) Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals. *Mol Biol Evol*, **25**, 709-718.

Isshiki, M., Tsumoto, A. and Shimamoto, K. (2006) The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. *Plant Cell*, **18**, 146-158.

James, A.B., Syed, N.H., Bordage, S., Marshall, J., Nimmo, G.A., Jenkins, G.I., et al. (2012) Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *Plant Cell*, **24**, 961-981.

Jeong, H.J., Kim, Y.J., Kim, S.H., Kim, Y.H., Lee, I.J., Kim, Y.K., et al. (2011) Nonsense-mediated mRNA decay factors, UPF1 and UPF3, contribute to plant defense. *Plant Cell Physiol*, **52**, 2147-2156.

Jurica, M.S. and Moore, M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell*, **12**, 5-14.

Kalyna, M., Lopato, S. and Barta, A. (2003) Ectopic expression of atRSZ33 reveals its function in splicing and causes pleiotropic changes in development. *Mol Biol Cell*, **14**, 3565-3577.

Kalyna, M., Lopato, S., Voronin, V. and Barta, A. (2006) Evolutionary conservation and regulation of particular alternative splicing events in plant SR proteins. *Nucleic Acids Res*, **34**, 4395-4405.

Kalyna, M., Simpson, C.G., Syed, N.H., Lewandowska, D., Marquez, Y., Kusenda, B., et al. (2012) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res*, **40**, 2454-2469.

Kazan, K. and Manners, J.M. (2008) Jasmonate signaling: toward an integrated view. *Plant Physiol*, **146**, 1459-1468.

Kerenyi, Z., Merai, Z., Hiripi, L., Benkovics, A., Gyula, P., Lacomme, C., et al. (2008) Inter-kingdom conservation of mechanism of nonsense-mediated mRNA decay. *EMBO J*, **27**, 1585-1595.

Kervestin, S. and Jacobson, A. (2012) NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol*, **13**, 700-712.

Kessler, A., Halitschke, R. and Baldwin, I.T. (2004) Silencing the jasmonate cascade: induced plant defenses and insect populations. *Science*, **305**, 665-668.

Kim, E., Magen, A. and Ast, G. (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, **35**, 125-131.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S. and Ahringer, J. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet*, **41**, 376-381.

Koren, E., Lev-Maor, G. and Ast, G. (2007) The emergence of alternative 3' and 5' splice site exons from constitutive exons. *PLoS Comput Biol*, **3**, e95.

Lambert, M.J., Cochran, W.O., Wilde, B.M., Olsen, K.G. and Cooper, C.D. (2015) Evidence for widespread subfunctionalization of splice forms in vertebrate genomes. *Genome Res*, **25**, 624-632.

Lareau, L.F. and Brenner, S.E. (2015) Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. *Mol Biol Evol*, **32**, 1072-1079.

Lee, B.H., Kapoor, A., Zhu, J. and Zhu, J.K. (2006) STABILIZED1, a stress-upregulated nuclear protein, is required for pre-mRNA splicing, mRNA turnover, and stress tolerance in Arabidopsis. *Plant Cell*, **18**, 1736-1749.

Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., et al. (2007) The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet*, **3**, e203.

Leviatan, N., Alkan, N., Leshkowitz, D. and Fluhr, R. (2013) Genome-wide survey of cold stress regulated alternative splicing in *Arabidopsis thaliana* with tiling microarray. *PLoS One*, **8**, e66511.

Lewandowska, D., Simpson, C.G., Clark, G.P., Jennings, N.S., Barciszewska-Pacak, M., Lin, C.F., et al. (2004) Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell*, **16**, 1340-1352.

Li, D., Heiling, S., Baldwin, I.T. and Gaquerel, E. (2016) Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. *Proc Natl Acad Sci U S A*.

Li, J., Li, X., Guo, L., Lu, F., Feng, X., He, K., et al. (2006a) A subgroup of MYB transcription factor genes undergoes highly conserved alternative splicing in Arabidopsis and rice. *J Exp Bot*, **57**, 1263-1273.

Li, W., Lin, W.D., Ray, P., Lan, P. and Schmidt, W. (2013) Genome-wide detection of condition-sensitive alternative splicing in Arabidopsis roots. *Plant Physiol*, **162**, 1750-1763.

Li, Y., Bor, Y.C., Misawa, Y., Xue, Y., Rekosh, D. and Hammarskjold, M.L. (2006b) An intron with a constitutive transport element is retained in a Tap messenger RNA. *Nature*, **443**, 234-237.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., et al. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464-469.

Ling, Z., Zhou, W., Baldwin, I.T. and Xu, S. (2015) Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuata*. *Plant J*, **84**, 228-243.

Liu, H.X., Goodall, G.J., Kole, R. and Filipowicz, W. (1995) Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in *Nicotiana plumbaginifolia*. *EMBO J*, **14**, 377-388.

Liu, J., Sun, N., Liu, M., Liu, J., Du, B., Wang, X., et al. (2013) An autoregulatory loop controlling Arabidopsis HsfA2 expression: role of heat shock-induced alternative splicing. *Plant Physiol*, **162**, 512-521.

Lopato, S., Gattoni, R., Fabini, G., Stevenin, J. and Barta, A. (1999a) A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. *Plant Mol Biol*, **39**, 761-773.

Lopato, S., Kalyna, M., Dorner, S., Kobayashi, R., Krainer, A.R. and Barta, A. (1999b) atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes. *Genes Dev*, **13**, 987-1001.

Lorkovic, Z.J. (2009) Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends Plant Sci*, **14**, 229-236.

Lorkovic, Z.J., Wieczorek Kirk, D.A., Lambermon, M.H. and Filipowicz, W. (2000) Pre-mRNA splicing in higher plants. *Trends Plant Sci*, **5**, 160-167.

Machado, R.A., Ferrieri, A.P., Robert, C.A., Glauser, G., Kallenbach, M., Baldwin, I.T., et al. (2013) Leaf-herbivore attack reduces carbon reserves and regrowth from the roots via jasmonate and auxin signaling. *New Phytol*, **200**, 1234-1246.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res*, **22**, 1184-1195.

Mastrangelo, A.M., Marone, D., Laido, G., De Leonardis, A.M. and De Vita, P. (2012) Alternative splicing: enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci*, **185-186**, 40-49.

Matsukura, S., Mizoi, J., Yoshida, T., Todaka, D., Ito, Y., Maruyama, K., et al. (2010) Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes. *Mol Genet Genomics*, **283**, 185-196.

Mattick, J.S. (1994) Introns: evolution and function. *Curr Opin Genet Dev*, **4**, 823-831.

Merkin, J., Russell, C., Chen, P. and Burge, C.B. (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593-1599.

Meyer, M., Plass, M., Perez-Valle, J., Eyras, E. and Vilardell, J. (2011) Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol Cell*, **43**, 1033-1039.

Michael Weaver, L., Swiderski, M.R., Li, Y. and Jones, J.D. (2006) The *Arabidopsis thaliana* TIR-NB-LRR R-protein, RPP1A; protein localization and constitutive activation of defence by truncated alleles in tobacco and Arabidopsis. *Plant J*, **47**, 829-840.

Miriami, E., Margalit, H. and Sperling, R. (2003) Conserved sequence elements associated with exon skipping. *Nucleic Acids Res*, **31**, 1974-1983.

Mitra, S. and Baldwin, I.T. (2008) Independently silencing two photosynthetic proteins in *Nicotiana attenuata* has different effects on herbivore resistance. *Plant Physiol*, **148**, 1128-1138.

Monks, D.E., Aghoram, K., Courtney, P.D., DeWald, D.B. and Dewey, R.E. (2001) Hyperosmotic stress induces the rapid phosphorylation of a soybean phosphatidylinositol transfer protein homolog through activation of the protein kinases SPK1 and SPK2. *Plant Cell*, **13**, 1205-1219.

Nyiko, T., Sonkoly, B., Merai, Z., Benkovics, A.H. and Silhavy, D. (2009) Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol Biol*, **71**, 367-378.

Palusa, S.G., Ali, G.S. and Reddy, A.S. (2007) Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J*, **49**, 1091-1107.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **40**, 1413-1415.

Pertea, M., Mount, S.M. and Salzberg, S.L. (2007) A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. *BMC Bioinformatics*, **8**, 159.

Qin, F., Kakimoto, M., Sakuma, Y., Maruyama, K., Osakabe, Y., Tran, L.S., et al. (2007) Regulation and functional analysis of ZmDREB2A in response to drought and heat stresses in *Zea mays* L. *Plant J*, **50**, 54-69.

Rappsilber, J., Ryder, U., Lamond, A.I. and Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome. *Genome Res*, **12**, 1231-1245.

Reddy, A.S. (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol*, **58**, 267-294.

Reddy, A.S., Marquez, Y., Kalyna, M. and Barta, A. (2013) Complexity of the alternative splicing landscape in plants. *Plant Cell*, **25**, 3657-3683.

Reddy, A.S., Rogers, M.F., Richardson, D.N., Hamilton, M. and Ben-Hur, A. (2012) Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Front Plant Sci*, **3**, 18.

Richardson, D.N., Rogers, M.F., Labadorf, A., Ben-Hur, A., Guo, H., Paterson, A.H., et al. (2011) Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into sub-family classification and extent of alternative splicing. *PLoS One*, **6**, e24542.

Riehs, N., Akimcheva, S., Puizina, J., Bulankova, P., Idol, R.A., Siroky, J., et al. (2008) Arabidopsis SMG7 protein is required for exit from meiosis. *J Cell Sci*, **121**, 2208-2216.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J. and Seelig, G. (2015) Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, **163**, 698-711.

Satyawan, D., Kim, M.Y. and Lee, S.H. (2016) Stochastic alternative splicing is prevalent in mungbean (*Vigna radiata*). *Plant Biotechnol J*.

Schindler, S., Szafranski, K., Hiller, M., Ali, G.S., Palusa, S.G., Backofen, R., et al. (2008) Alternative splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes. *BMC Genomics*, **9**, 159.

Schoenberg, D.R. and Maquat, L.E. (2012) Regulation of cytoplasmic mRNA decay. *Nat Rev Genet*, **13**, 246-259.

Schoning, J.C., Streitner, C., Meyer, I.M., Gao, Y. and Staiger, D. (2008) Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis. *Nucleic Acids Res*, **36**, 6977-6987.

Schwachtje, J. and Baldwin, I.T. (2008) Why does herbivore attack reconfigure primary metabolism? *Plant Physiol*, **146**, 845-851.

Schwachtje, J., Minchin, P.E., Jahnke, S., van Dongen, J.T., Schittko, U. and Baldwin, I.T. (2006) SNF1-related kinases allow plants to tolerate herbivory by allocating carbon to roots. *Proc Natl Acad Sci U S A*, **103**, 12935-12940.

Schwartz, S., Meshorer, E. and Ast, G. (2009) Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, **16**, 990-995.

Seo, P.J., Kim, M.J., Ryu, J.Y., Jeong, E.Y. and Park, C.M. (2011) Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. *Nat Commun*, **2**, 303.

Seo, P.J., Park, M.J., Lim, M.H., Kim, S.G., Lee, M., Baldwin, I.T., et al. (2012) A self-regulatory circuit of CIRCADIAN CLOCK-ASSOCIATED1 underlies the circadian clock regulation of temperature responses in Arabidopsis. *Plant Cell*, **24**, 2427-2442.

Severing, E.I., van Dijk, A.D., Stiekema, W.J. and van Ham, R.C. (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics*, **10**, 154.

Shen, Y., Zhou, Z., Wang, Z., Li, W., Fang, C., Wu, M., et al. (2014) Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell*, **26**, 996-1008.

Sierro, N., Battey, J.N., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., et al. (2014) The tobacco genome sequence and its comparison with those of tomato and potato. *Nat Commun*, **5**, 3833.

Silverman, I.M., Li, F. and Gregory, B.D. (2013) Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. *Plant Sci*, **205-206**, 55-62.

Simpson, C.G., Fuller, J., Maronova, M., Kalyna, M., Davidson, D., McNicol, J., et al. (2008) Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant J*, **53**, 1035-1048.

Simpson, C.G., Jennings, S.N., Clark, G.P., Thow, G. and Brown, J.W. (2004) Dual functionality of a plant U-rich intronic sequence element. *Plant J*, **37**, 82-91.

Slotte, T., Hazzouri, K.M., Agren, J.A., Koenig, D., Maumus, F., Guo, Y.L., et al. (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet*, **45**, 831-835.

Smith, C.W., Chu, T.T. and Nadal-Ginard, B. (1993) Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol*, **13**, 4939-4952.

Smith, C.W., Porro, E.B., Patton, J.G. and Nadal-Ginard, B. (1989) Scanning from an independently specified branch point defines the 3' splice site of mammalian introns. *Nature*, **342**, 243-247.

Sorek, R., Ast, G. and Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res*, **12**, 1060-1067.

Staiger, D. and Brown, J.W. (2013) Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell*, **25**, 3640-3656.

Staudt, A.C. and Wenkel, S. (2011) Regulation of protein function by 'microProteins'. *EMBO Rep*, **12**, 35-42.

Stauffer, E., Westermann, A., Wagner, G. and Wachter, A. (2010) Polypyrimidine tract-binding protein homologues from Arabidopsis underlie regulatory circuits based on alternative splicing and downstream control. *Plant J*, **64**, 243-255.

Steppuhn, A., Gase, K., Krock, B., Halitschke, R. and Baldwin, I.T. (2004) Nicotine's defensive function in nature. *PLoS Biol*, **2**, E217.

Sterner, D.A., Carlo, T. and Berget, S.M. (1996) Architectural limits on split genes. *Proc Natl Acad Sci U S A*, **93**, 15081-15085.

Streitner, C., Simpson, C.G., Shaw, P., Danisman, S., Brown, J.W. and Staiger, D. (2013) Small changes in ambient temperature affect alternative splicing in *Arabidopsis thaliana*. *Plant Signal Behav*, **8**, e24638.

Su, Z., Wang, J., Yu, J., Huang, X. and Gu, X. (2006) Evolution of alternative splicing after gene duplication. *Genome Res*, **16**, 182-189.

Sugio, A., Dreos, R., Aparicio, F. and Maule, A.J. (2009) The cytosolic protein response as a subcomponent of the wider heat shock response in Arabidopsis. *Plant Cell*, **21**, 642-654.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A. and Brown, J.W. (2012) Alternative splicing in plants--coming of age. *Trends Plant Sci*, **17**, 616-623.

Syed, N.H., Prince, S.J., Mutava, R.N., Patil, G., Li, S., Chen, W., et al. (2015) Core clock, SUB1, and ABAR genes mediate flooding and drought responses via alternative splicing in soybean. *J Exp Bot*, **66**, 7129-7149.

Talerico, M. and Berget, S.M. (1994) Intron definition in splicing of small Drosophila introns. *Mol Cell Biol*, **14**, 3434-3445.

Tanabe, N., Yoshimura, K., Kimura, A., Yabuta, Y. and Shigeoka, S. (2007) Differential expression of alternatively spliced mRNAs of Arabidopsis SR protein homologs, atSR30 and atSR45a, in response to environmental stress. *Plant Cell Physiol*, **48**, 1036-1049.

Thatcher, S.R., Danilevskaya, O.N., Meng, X., Beatty, M., Zastrow-Hayes, G., Harris, C., et al. (2016) Genome-Wide Analysis of Alternative Splicing during Development and Drought Stress in Maize. *Plant Physiol*, **170**, 586-599.

Thomas, J., Palusa, S.G., Prasad, K.V., Ali, G.S., Surabhi, G.K., Ben-Hur, A., et al. (2012) Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of SCL33 pre-mRNA. *Plant J*, **72**, 935-946.

Tillemans, V., Dispa, L., Remacle, C., Collinge, M. and Motte, P. (2005) Functional distribution and dynamics of Arabidopsis SR splicing factors in living plant cells. *Plant J*, **41**, 567-582.

Tillemans, V., Leponce, I., Rausin, G., Dispa, L. and Motte, P. (2006) Insights into nuclear organization in plants as revealed by the dynamic distribution of Arabidopsis SR splicing factors. *Plant Cell*, **18**, 3218-3234.

Tomato Genome, C. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635-641.

Vogan, K.J., Underhill, D.A. and Gros, P. (1996) An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. *Mol Cell Biol*, **16**, 6677-6686.

Wachter, A., Ruhl, C. and Stauffer, E. (2012) The Role of Polypyrimidine Tract-Binding Proteins and Other hnRNP Proteins in Plant Splicing Regulation. *Front Plant Sci*, **3**, 81.

Wang, B.B. and Brendel, V. (2006) Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci U S A*, **103**, 7175-7180.

Wang, B.B., O'Toole, M., Brendel, V. and Young, N.D. (2008) Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. *BMC Plant Biol*, **8**, 17.

Warf, M.B. and Berglund, J.A. (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci*, **35**, 169-178.

Werneke, J.M., Chatfield, J.M. and Ogren, W.L. (1989) Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and Arabidopsis. *Plant Cell*, **1**, 815-825.

Wu, G. and Poethig, R.S. (2006) Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3. *Development*, **133**, 3539-3547.

Wu, H.P., Su, Y.S., Chen, H.C., Chen, Y.R., Wu, C.C., Lin, W.D., et al. (2014) Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in *Physcomitrella patens*. *Genome Biol*, **15**, R10.

Wu, J. and Baldwin, I.T. (2010) New insights into plant responses to the attack from insect herbivores. *Annu Rev Genet*, **44**, 1-24.

Wu, J., Hettenhausen, C., Meldau, S. and Baldwin, I.T. (2007a) Herbivory rapidly activates MAPK signaling in attacked and unattacked leaf regions but not between leaves of *Nicotiana attenuata*. *Plant Cell*, **19**, 1096-1122.

Wu, J., Kang, J.H., Hettenhausen, C. and Baldwin, I.T. (2007b) Nonsense-mediated mRNA decay (NMD) silences the accumulation of aberrant trypsin proteinase inhibitor mRNA in *Nicotiana attenuata*. *Plant J*, **51**, 693-706.

Xiong, L., Schumaker, K.S. and Zhu, J.K. (2002) Cell signaling during cold, drought, and salt stress. *Plant Cell*, **14 Suppl**, S165-183.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (2006) Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu Rev Plant Biol*, **57**, 781-803.

Yan, K., Liu, P., Wu, C.A., Yang, G.D., Xu, R., Guo, Q.H., et al. (2012) Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in *Arabidopsis thaliana*. *Mol Cell*, **48**, 521-531.

Yang, G.D., Yan, K., Wu, B.J., Wang, Y.H., Gao, Y.X. and Zheng, C.C. (2012a) Genomewide analysis of intronic microRNAs in rice and Arabidopsis. *J Genet*, **91**, 313-324.

Yang, R. and Wang, X. (2013) Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *Plant Cell*, **25**, 71-82.

Yang, X., Zhang, H. and Li, L. (2012b) Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. *Plant J*, **70**, 421-431.

Yoshimura, K., Yabuta, Y., Ishikawa, T. and Shigeoka, S. (2002) Identification of a cis element for tissue-specific alternative splicing of chloroplast ascorbate peroxidase pre-mRNA in higher plants. *J Biol Chem*, **277**, 40623-40632.

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., et al. (2010) Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq. *Plant Physiol*, **152**, 1787-1795.

Zhang, F., Zhu, G., Du, L., Shang, X., Cheng, C., Yang, B., et al. (2016) Genetic regulation of salt stress tolerance revealed by RNA-Seq in cotton diploid wild species, *Gossypium davidsonii*. *Sci Rep*, **6**, 20582.

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., et al. (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*, **20**, 646-654.

Zhang, X.C. and Gassmann, W. (2003) RPS4-mediated disease resistance requires the combined presence of RPS4 transcripts with full-length and truncated open reading frames. *Plant Cell*, **15**, 2333-2342.

Zhang, X.C. and Gassmann, W. (2007) Alternative splicing and mRNA levels of the disease resistance gene RPS4 are induced during defense responses. *Plant Physiol*, **145**, 1577-1587.

Zhang, X.N. and Mount, S.M. (2009) Two alternatively spliced isoforms of the Arabidopsis SR45 protein have distinct roles during normal plant development. *Plant Physiol*, **150**, 1450-1458.

Eigenständigkeitserklärung

Entsprechend der geltenden, mir bekannten Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena erkläre ich, daß ich die vorliegende Dissertation eigenständig angefertigt und alle von mir benutzten Hilfsmittel und Quellen angegeben habe. Personen, die mich bei der Auswahl und Auswertung des Materials sowie bei der Fertigstellung der Manuskripte unterstützt haben, sind am Beginn eines jeden Kapitels genannt. Es wurde weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte für Arbeiten, welche im Zusammenhang mit dem Inhalt der vorliegenden Dissertation stehen, geldwerte Leistungen erhalten. Die vorgelegte Dissertation wurde außerdem weder als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung noch als Dissertation an einer anderen Hochschule eingereicht.

_____

Zhihao Ling

Jena, 06, 2017

Curriculum vitae

Ling Zhihao

Personal details

Date of Birth: 10.08.1988 Email: lzhdennisdn@gmail.com

Nationality: China Gender: Male

Education

PhD study
Max Planck Institute for Chemical Ecology (MPI-CE) Jena, Germany
(09, 2012 ~ Present)

Master of Natural Science in Bioinformatics
Lund University Lund, Sweden
(09, 2010 ~ 09, 2012)

Bachelor of Bioscience
Northwest A&F University Yangling, China
(09, 2006 ~ 09, 2010)

Research experience

09, 2012 ~ Present Max Planck Institute for Chemical Ecology (MPI-CE) Jena, Germany

PhD student (financed by MPI stipend)

- Analysing the genome of *Nicotiana attenuate*.
- Discovering insect herbivore induced alternative splicing (AS) in plant (based on NGS transcriptome data).
- Investigating the evolution of AS in eudicots.
- Developing pipeline for promoter analysis.
- Presenting results in international conference to get feedback and find cooperation.
- Publishing scientific results in peer-reviewed journals (two as first author).

11, 2011 ~ 12, 2011 Centre for Genomic Regulation (CRG) Barcelona, Spain

Exchange visit student (financed by Lund University)

- Studying basic bioinformatics analysis, programming and analysing fungal genome.

01, 2012 ~ 09, 2012 Lund University Lund, Sweden

Master student

- Analysing the genome of yeast *Dekkera bruxellensis*.
- Joining the development of an automatic genome annotation pipeline for new yeast strain.

Further education

"Hunting for Promoters" Bioinformatics Training Course Oeiras, Portugal

(10, 2012)

Univariate Statistics in Ecology and Evolution Summer school Bremen, Germany

(06, 2014)

Teaching experience

Tutoring and supervising exercise in graduate course in insect chemical ecology workshop for RNA-seq, Max Planck Institute for Chemical Ecology, Germany *(06, 2016).*

Tutoring and supervising exercise in the workshop of comparative genomics, University of Novo gorica, Slovenia *(07, 2012).*

Tutoring and supervising exercise in Molecular Biotechnology, Lund University, Sweden *(05, 2012)*

Skills

Bioinformatics: RNA-seq, Genome and transcriptome analysis

Molecular biology: PCR, RT-qPCR

IT: Python, R, Adobe illustrator Computer environment: Linux/Windows

Language

Chinese – mother tongue English - fluent

Publications

1. **Ling Z**., Brockmöller T., Baldwin, I. T., Xu S. (2016). The rapid evolution of alternative splicing in plants. (In revision to Plant physiology)
2. Zhou, W., Brockmöller, T., **Ling, Z**., Omdahl, A., Baldwin, I. T., Xu, S. (2016). Evolution of herbivore associated elicitor induced early defence signalling networks was shaped by genome-wide duplications in *Nicotiana*. *eLife* DOI: 10.7554/eLife.19531
3. Brockmöller, T., **Ling, Z**., Li, D., Gaquerel, E., Baldwin, I.T., Xu, S. (2016). *Nicotiana attenuata* Data Hub (*Na*DH): an integrative platform for exploring genomic, tran-scriptomic and metabolomic data in wild tobacco. *BMC Genomics*
4. Xu, S., Brockmöller, T., Navarro A., Gase K., **Ling, Z**., Stanke, M., Tang, Lyons, E., Kuhl, H., Timmermann, B., Gaquerel, E., Baldwin, I. T. (2016). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *PNAS*
5. Ishchuk, O.P., Vojvoda Zeljko, T., Schifferdecker, A.J., Mebrahtu Wisen, S., Hagstrom, A.K., Rozpedowska, E., Rordam Andersen, M., Hellborg, L., **Ling, Z**., Sibirny, A.A., and Piskur, J. (2016). Novel Centromeric Loci of the Wine and Beer Yeast *Dekkera bruxellensis* CEN1 and CEN2. *PloS one* 11, e0161741.
6. Schifferdecker, A. J., Siurkas, J., Andersen, M. R., Joerck-Ramberg, D., **Ling, Z**., Zhou, N., Blevins, J. E., Sibirny, A. A., Piškur, J., Ishchuk, O. P. (2016). Alcohol dehydrogenase gene ADH3 activates glucose alcoholic fermentation in genetically engineered *Dekkera bruxellensis* yeast. *Applied Microbiology and Biotechnology* 100, 3219-3231
7. **Ling Z**, Zhou W, Baldwin, I. T, Xu S (2015). Insect herbivory elicits genome-wide alternative splicing responses in *Nicotiana attenuate*. *The Plant Journal* 84**,** 228-243
8. Moktaduzzaman M, Galafassi S, Vigentini I, **Ling Z**, Piškur J, Compagno C (2015) Galactose utilization sheds new light on sugar metabolism in the sequenced strain *Dekkera bruxellensis* CBS 2499. *FEMS Yeast Research* 15
9. Piškur J, **Ling Z**, Marcet-Houben M, Ishchuk OP, Aerts A, LaButti K, Copeland A, Lindquist E, Barry K, Compagno C, Bisson L, Grigoriev IV, Gabaldón T, Phister T (2012) The genome of wine yeast *Dekkera bruxellensis*. *International Journal of Food Microbiol* 157, 202-209

Oral Presentations

Ling Z. (2015). Insect herbivore elicits genome-wide alternative splicing responses in *Nicotiana attenuata*. Talk presented at 14th IMPRS Symposium, MPI for Chemical Ecology, Dornburg, DE

Poster Presentations

Guo H., Bhattacharya S., Diezel C., **Ling Z**., Xu S., Wielsch N., Svatoš A., Baldwin I.T. (2015). Molecular analysis of S-RNase genes in mate selection of *N. attenuata*: self-incompatibility genes still work in self-compatible species. Poster presented at ICE Symposium, MPI for Chemical Ecology, Jena, DE

Ling Z., Brockmöller T., Baldwin I.T., Xu S. (2015). The evolution of alternative splicing and gene expression among closely related *Nicotiana* species. Poster presented at Annual meeting of the Society for Molecular Biology and Evolution 2015, Society for Molecular Biology and Evolution, vienna, AT

Xu S., Gaquerel E., Navarro-Quezada A., Brockmöller T., Gase K., **Ling Z**., Tang H., Lyons E., Kuhl H., Timmermann B., Baldwin I.T. (2014). The *Nicotiana* genome projects. Poster presented at SAB Meeting 2014, MPI for Chemical Ecology, Jena, DE

Ling Z., Baldwin I.T., Xu S. (2014). Herbivore induced Alternative Splicing in *Nicotiana attenuata*. Poster presented at 13th IMPRS Symposium, MPI for Chemical Ecology, Dornburg, DE

Ling Z., Baldwin I.T., Xu S. (2013). Transcriptome survey on the alternative splicing landscape in *Nicotiana attenuata*. Poster presented at Plant Genome Evolution Conference, Amsterdam, NL

Ling Z. (2013). Transcriptome survey on the alternative splicing landscape in *Nicotiana attenuata*. Poster presented at 12th IMPRS Symposium, MPI for Chemical Ecology, Jena, DE

Acknowledgement

Finally it is time to thank everyone who gave me different kinds of support during my time as a Ph.D. student. A big thank you to all of you, without you, it is impossible for me to accomplish this work.

I would like express my deepest gratitude to Prof. **Ian T. Baldwin**, who gave me the opportunity to pursue my Ph.D degree in this excellent department. Your passion in science inspired me throughout all the years. Thank you for your patient guidance, I learned a lot from you, especially for how to think as a scientist.

My hearty thanks to my supervisor Dr. **Shuqing Xu**, who always provide me support not only on my work but also on my daily life. I could not thank you enough for taking me as your first Ph.D. student. I really feel calm under your supervision, because you can always quickly find a solution when I got lost in my research. Thank you for training my writing skill and encouraging me to go to conference.

Thanks to Prof. **Ralf Oelmüller** for scientific discussion and be a member of my committee meeting.

Thanks to my colleagues at the Max Planck Institute for Chemical Ecology for all the helps. I thank my group members **Thomas Brockmöller** and **Wenwu Zhou** for fruitful scientific discussion and participate in my projects; **Henrique Valim** for proofreading and correcting my grammar mistakes of my thesis; **Sven Heiling** for helping me with the Zusammenfassung of my thesis; **Klaus Gase** for helping me upload the data; **Martin Niebergall** for helping me solve the computer problems; **Evelyn Clauβen** for her organizational efforts. My big thanks to **Rakesh Santhanam**, **Lucas Cortés Llorca**, **Youngsung Joo**, **Van Thi Luu**, **Martin Schäfer**, **Meredith C. Schuman**, **Christoph Brütting** and **Felipe Yon** for scientific and other discussion. I enjoyed talking to you a lot. My big thanks would next go to our Chinese community, **Dapeng Li**, **Han Guo**, **Dechang Cao**, **Ming Wang**, **Xiang Li**, **Ran Li**, **Jiancai Li**, **Yang Wang**, **Jun He**, **YuanYuan Zhao** and **Zhiling Yang**. You have made my work more enjoyable and comfortable. Thanks to all the scientific and non-scientific discussions in the lab, during lunch and coffee break. I would also like to thank **my basketball team** in Jena, you made me fall in love with the sport and kept me looking forward to every Saturday.

My deep thanks to my master supervisor **Prof. Jure Piskur**, who paved the way for me to science. Without him, my life would be completely different. R.I.P my dear professor, you will live in my heart forever.

Last but not least, I want to thank my beloved family really a lot for all their supports, your constantly encouragements always give my power to face every problems confidently and optimistically.