ABSTRACT

DLUGE, KURTIS LAWRENCE. Application of Genomic Approaches Toward Characterization of Commercially Important Tobacco Loci. (Under the direction of Dr. Ralph Dewey.)

Advancements in next generation sequencing (NGS) and the emergence of new tobacco genetic resources are allowing new investigations into specific chromosomal regions that harbor traits of agronomic interest. Here, the production of expression data using RNA-seq technologies was used in conjunction with recent tobacco draft genomes and annotations to characterize three tobacco loci of commercial importance. The *VAM* locus represents a large deletion mutation that provides resistance to certain Potyvirus strains, as well as mediating severe reductions in the levels of key trichome exudates. The locus designated *va*, likely derived from *VAM*, gives a similar Potyvirus resistance phenotype without the associated trichome exudate defects. Using RNA-seq to characterize the transcriptomes of tobacco cultivars K326, K326-*va*, and TI 1406 (a source of *VAM*), we produced a list of contigs expected to be missing in genomes possessing the *VAM* or *va* deletion mutations. Alignment of candidate contigs to anchored genomic scaffolds provided a framework for defining the chromosomal architecture of the VAM/*va* loci. Through the utilization of publically available genetic markers and genome assemblies from collaborators, we were able to bridge gaps between anchored scaffolds and provide the most comprehensive characterization of these deletions to date. In total, we found that the *VAM* deletion spans scaffolds encompassing 6.55Mbp, with ~5.5Mbp of this also corresponding to the *va* mutation. The region defined by *VAM* contains a minimum of 184 genes. Dominant PCR-based markers very near one of the *va* deletion junctions were also developed. Transformation experiments in *VAM* and *va* cultivars also verified that restoring

eiF4E.S1, a pivotal gene that is missing in both *VAM* and *va* mutations, is sufficient to restore susceptibility to potato virus Y (PVY) and tobacco etch virus (TEV). Additionally, it was found that among the genes missing in plants possessing *VAM,* but not *va,* were members of the cembratrienol synthase gene family (CBTS), whose activity is required for production of the major leaf surface diterpenes α-and β-cembratrien-diols. Interestingly, transgenic TI 1406 (*VAM*) plants transformed with a CBTS gene still failed to produce the missing trichrome diterpene compounds, presumably due to the absence of an additional gene(s) within *VAM* that is responsible for the production of functional chloroplasts within the trichome heads.

A similar strategy was also used to characterize a third chromosomal region of interest; specifically, a large introgression fragment from the wild *species N. debneyi* that confers resistance to *Peronospora tabacina* (blue mold). Here, the blue mold resistance (*bmr*) introgression fragment was investigated through RNA-seq based analysis of line NC 775 bmr/bmr, a near-isogenic derivative of burley cultivar NC 775 which obtained the *bmr* locus through backcrossing to the blue mold resistant variety NC 2000. Through sequencing, assembly and alignment, we were able to identify and validate 26 contigs that are located on the *bmr* introgression fragment. These were subsequently mapped across the introgression region through the use of 14 recombinant lines that we developed. Field evaluation for blue mold resistance of the parental line NC 775 bmr/bmr lead to the unexpected observation that despite receiving the *N. debneyi*-derived introgression fragment from NC 2000 (using two SCAR markers specific to that region during the backcrossing process), NC 2000 displayed substantially great resistance to the blue mold

pathogen. This result prompted an RNA-seq-based analysis of the NC 2000 variety as well, to gain insights into the source of the additional blue resistance QTL (quantitative trait loci) found in NC 2000. Finally, another unexpected result was the discovery that the *bmr* introgression fragment was also highly associated with enhanced susceptibility to the disease black shank within the NC 775 background.

Genomic Approaches for, and Characterization of Commercially Important Tobacco Loci

by
Kurtis Lawrence Dluge

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Functional Genomics

Raleigh, North Carolina

2017

APPROVED BY:

_____	_____
Dr. Ralph E. Dewey Committee Chair	Dr. Ramsey S. Lewis
_____	_____
Dr. Ross W. Whetten	Dr. Jean B. Ristaino

BIOGRAPHY

Kurtis Dluge started in Michigan. Went to WMU. Went to NCSU.

“Kurtis is a real fly sorta guy” -Abraham Lincoln

ACKNOWLEDGMENTS

I would like to first thank my advisor, Dr. Ralph Dewey for his guidance and support opportunities he has provided me. I could not have asked for a better advisor. Thanks to all the helpful and friendly past and present members of the Dewey lab; Carol Griffin, Jianli Lu, William Smith, Christophe La Hovary, and others. I would also like to express gratitude to the other members of my graduate committee: Dr. Ramsey Lewis, Dr. Ross Whetten, and Dr. Jean Ristaino for their direction and assistance. An additional thanks is also due to British American Tobacco for their financial support and the Yunnan Academy of Tobacco Agricultural Sciences (YATAS) for their financial support and assistance with projects.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Literature Review

Uncharacterized and potentially underutilized genetic variability is assumed to exist within the germplasm collections of crop species (Gur and Zamir, 2004). In terms of availability and number of germplasm collections and breeding lines, the 76 species genus *Nicotiana* does not fall short. Approximately 1900 accessions of the commercially important *Nicotiana tabacum* are available to researchers and breeders from the U.S. *Nicotiana* germplasm collection, with many additional collections and species held outside of the U.S. (Lewis and Nicholson, 2007). Among other things, tobacco accessions found in these collections have been used to study the relatedness among tobaccos (Lewis and Nicholson, 2007; Fricano et al., 2012) and to probe for the presence of genes linked to disease resistance (Julio et al., 2015). In the large and complex genome of tobacco, recent next-generation sequencing (NGS) techniques and updated genomic resources are allowing better and deeper investigation into the characteristics that make up useful genetic variability found in these collections. Characterizing genes and developing markers associated with important agronomic traits can proceed at a faster pace than was previously possible.

Current tobacco genome and transcriptome NGS strategies and technologies.

Nicotiana tabacum, an allotetraploid (2n =4x =48), has a genome size of approximately 4.5Gb (Edwards et al., 2017). Its genome arose from the hybridization of the two ancestral parents, *N. sylvestris* (maternal S genome) and *N. tomentosiformis* (paternal T genome), each contributing one set of each of their chromosomes (Kenton et al., 1993; Sierro et al., 2014). The *N. tabacum* genome is roughly 85% of the size of the two ancestral genomes combined, with most of the loss corresponding to T-genome repetitive sequences (Edwards et al., 2017; Bennett and Leitch, 1995). Genome sequencing has also shown that it is unlikely that any other *Nicotiana* species' DNA is represented in *N. tabacum* (Sierro et al., 2014; Edwards et al., 2017). The *N. tabacum* genome encodes around 69,500 genes, many more than the predicted 27,416 genes of *A. thaliana* (Arabidopsis) or the 35,119 genes of *S. tuberosum* (potato) (Edwards et al., 2017). Repeat families are expected to cover ~67% of the *N. tabacum* genome (Edwards et al., 2017). This large genome size and abundance of repetitive elements adds to the difficulty of performing NGS in tobacco in comparison to many other plant species.

NGS has been commonly used in tobacco when measuring RNA expression at a transcriptome level and in assembling genomes (Edwards et al., 2017; Sierro et al., 2014). For these applications, the most important considerations (in addition to cost) when choosing a NGS platform are read length, read accuracy, and number of reads attainable. To a point, a higher number of NGS reads results in more accurate assemblies and transcriptome-based expression measurements (Cai et al., 2017). In general, assembly of

genomes and the measurement of gene expression using NGS is dependent on aligning reads to each other (*de novo*) or to a reference genome. Therefore, the longer the reads the less likely they are to be misaligned. For *N. tabacum*, its two ancestral genomes are highly similar (Sierro et al., 2013), so longer reads may be the only way to distinguish gene families with multiple closely related isoforms. Additionally, read accuracy issues can often be solved with greater read depth.

Currently, the most commonly used NGS technology is sequencing by synthesis (SBS). Illumina's most popular platforms utilize this sequencing method (Goodwin et al., 2016). An experiment designed to run on an Illumina SBS-based instrument begins with the addition of adapter sequences to the ends of a genomic or cDNA library that has been fragmented and selected for a desired length. Adapters can include nucleotide signatures called barcodes, enabling the multiplexing of experiments in a single run. The sequencing and sorting by barcode is then used to distinguish experimental sets. The adapter ends of the single-strand library fragments are then hybridized to a flow cell, which is a "slide" covered in bound complimentary adapters. Each individually attached strand is then duplicated through a process called solid-phase bridge amplification. This is when the free end of the attached strand bends over and attaches to the chip, followed by complementary strand synthesis. Upon DNA denaturation, one is left with two strands attached at their bases to the flow cell. Several rounds of this synthesis and denaturation will produce a clonal cluster of the same strand in one area of the flow cell. The flow cell is then ready for sequencing.

Four uniquely fluorescent, fluorophore-labeled nucleotides, blocked on their terminal end, are washed over the flow cell and hybridized to their complementary base (Guo et al., 2008). The slide is then imaged with each clonal cluster giving one of four fluorescent signals. The fluorophore is then cleaved and washed away. The terminal end of the nucleotide is restored and the cycle repeated for the next nucleotide in line on the strand. For paired-end (PE) reads, the strand can be sequenced in the same way from the opposite end as well, potentially doubling the sequence reads per strand. Often, PE reads are designed so that sequencing from both ends results in a single final product. However, a gap of unsequenced reads can be left which can provide unique sequencing information. The Illumina MiSeq v3 can potentially produce 50M 300bp PE reads (600bp read per strand) per run and the Illumina HiSeq2500 v4 can potentially produce 4B 125bp PE reads per run (Goodwin et al., 2016).

Previously, Roche's 454 pyrosequencing technology had been used in *Nicotiana* genome sequencing projects (Sierro et al., 2013). An advantage of 454 sequencing is that it can produce reads up to 1000bp long. However, 454 technology proved to be inferior compared to subsequently developed technologies and is no longer supported by Roche. Currently, the most commonly used long read NGS platforms use a single-molecule real-time (SMRT) sequencing approach (Goodwin et al., 2016). SMRT technologies use different techniques to measure polymerase activity as it synthesizes a single complementary DNA strand. The most common platform using this technology, Pacific Biosciences instruments, measures the light emitted when one of four labeled dNTPs is incorporated into an

anchored polymerase within a well containing only one nucleotide strand (Eid et al., 2009). For a single run on a Pacific Biosciences Sequel, ~350,000 reads of up to 12kb in length can be produced (Goodwin et al., 2016). However, error rates of >10% are common. To mitigate these errors, the circularized state of the target strand during sequencing can be harnessed, meaning multiple reads of each nucleotide can be made as the polymerase cycles around the circular template. As errors are generally random, a circular consensus read can be produced with a lower overall error rate (Goodwin et al., 2016).

While circular consensus reads can be used to mitigate errors in SMRT sequencing, in Illumina SBS, whose read accuracy for certain instruments is currently >99.5%, errors (substitution errors being the most common) are most frequently addressed by sequencing to a greater depth using quality control and trimming programs post-sequencing (Bolger et al., 2014; Goodwin et al., 2016).

Current tobacco genomic resources

Though *N. tabacum* is a very important and widely used model organism, until 2014 the lack of a draft genome had hindered certain avenues of experimentation. Draft genomes within the *Nicotiana* genus now exist of varying quality for *N. sylvestris* and *N. tomentosiformis* (Sierro et al., 2013), *N. attenuata* (Xu et al., 2017), *N. obtusifolia* (Xu et al., 2017), *N. benthamiana* (Bombarely et al., 2012), *N. otophora* (Sierro et al., 2014), and three cultivars of *N. tabacum*: K326 (Sierro et al., 2014**;** Edwards et al., 2017), TN90 (Sierro et al., 2014), and Basma Xanthi (BX) (Sierro et al., 2014). In general, the most important

assemblies to researchers and the tobacco industry alike are those of the *N. tabacum* cultivars, as this is the dominant species from both a research and commercial perspective. The majority of the *N. tabacum* planted in the U.S is of the Burley (ex: TN90) or Flue-cured (ex: K326) market types. When Fricano et al. (2012) looked at molecular markers in *N. tabacum* cultivars, they showed that markers within each of the Burley, Flue-cured, and Oriental types were more homogenous than those of Cigar, Primitive, Dark, or other tobacco classes, which showed more heterogeneous marker groupings. The differences, especially phenotypically, among these classes are important commercially, but in general there is a high degree of genetic relatedness across all commercial tobacco varieties (Murphy et al., 1987). Thus, the three sequenced *N. tabacum* cultivars should be generally useful for studies in any *N. tabacum* cultivar or market type.

A brief general review of *de novo* genome construction can be beneficial to understand the differences between related draft genomes. First, the ends of processed PE or single-end (SE) reads are aligned with each other to form contigs. Mate-pair (MP) reads, which are separated by a much larger distance than PE reads between the sequenced ends, are then used to group contigs into scaffolds. Finally, multiple methods, including the use of Bacterial Artificial Chromosome (BAC) libraries, optical mapping, and markers, can be used to join scaffolds or place them on a chromosome.

The first *N. tabacum* K326 draft genome (along with TN90 and BX) was released by Sierro et al. (2014) (referred to here as the PMI genome). It was constructed using ~710M Illumina HiSeq-2000 100bp PE or MP reads representing a 38X sequencing coverage. This

resulted in 582,565 scaffolds with an estimated genome coverage of 81.1% (Sierro et al., 2014). Edwards et al. (2017) later released their own K326 draft genome (referred to here as the BAT genome) constructed using ~192M Roche 454 and ~2.47B Illumina HiSeq-2000 PE and MP reads representing a combined 86X sequencing coverage. This resulted in 440,772 scaffolds with an estimated genome coverage of 90% (Edwards et al., 2017). Genomes assembled using multiple NGS technologies, such as the BAT genome, may be more reliable as they can partially avoid the errors inherent in any individual sequencing technology (Goodwin et al., 2016). During draft construction, after initial scaffolding using MP reads, both K326 assemblies utilized Whole Genome Profiling (WGP) to further extend scaffolds. Briefly, this is performed when a BAC library representing the species is sequenced (in these cases using NGS) at restriction digestion landmarks to generate sequence tags. A WGP map is then constructed by aligning the sequence tags. Current scaffolds can then be expanded or connected by aligning with the WGP map using its sequence tags (Sierro et al., 2013b). The chromosomes of both assemblies were ordered to agree with the genetic map of Bindler et al. (2011).

Subsequent construction of the two *N. tabacum* K326 draft genomes differed, as the PMI genome utilized MP libraries from the *N. tabacum* ancestral parental genomes *N. sylvestris* and *N. tomentosiformis* (Sierro et al., 2013) to improve scaffolding, whereas the BAT genome utilized an optical map to further improve its scaffolds, and for anchoring. An anchored contig/scaffold is one in which a chromosomal location has been assigned to a physical map (Mascher and Stein, 2014). Anchoring does not necessarily mean that the

scaffold/contig orientation, the precise distance from chromosome end to scaffold/contig, or exact the distance between scaffold/contigs are known, but ordering of scaffolds should be correct (Edwards et al., 2017). Optical mapping is performed by producing high molecular weight (HMW) genomic DNA (~1Mb) and labeling it with a nickase-attached fluorescent marker. The imaging and aligning of these results in a map of very large chromosomal sections, much like WGP maps. Scaffolds that align with this map can then be anchored to a chromosome (Goodwin et al., 2016). The BAT draft further anchored scaffolds using the 2,015 *N. tabacum* Infinium-30K HD consensus map's single nucleotide polymorphism (SNP) genetic markers (https://solgenomics.net/cview/map.pl?map_version_id=178) (Edwards et al., 2017). It appears that markers that did not fully align to a scaffold, or gave indistinct mapping results were not included in the final assembly.

Although the BAT assembly has a larger overall genome coverage and lower levels of missing or fragmented gene sequences compared to the PMI assembly, the differences in their construction means both can be useful and certain gene families may be better represented in one over the other. In addition to the other *Nicotiana* genome assemblies, the closely related species potato (*Solanum tuberosum*) and tomato (*Solanum lycopersicum*) have well characterized genomes and are useful tools for those working in tobacco. Tobacco gene expression microarrays have also been developed and can be useful in specialized situations. In general, expression microarrays are less robust in distinguishing closely related genes due to the short (compared to NGS technology) probes used to bind RNA. This is of

extra concern in the allotetraploid tobacco where a high percentage of its genes display redundancies. For many applications, the transcriptomes released along with the BAT and PMI genome assemblies are complete enough for expression analysis and alignment for studies using RNA-seq technologies.

Disease resistance/tolerance and management in tobacco

In North Carolina, diseases in tobacco from such diverse sources as nematodes, viruses, and oomycetes, can cause tens of millions of dollars in loss in any given year due to reductions in yield and quality, and can completely devastate individual fields (Mila and Radcliff, 2010). In *Nicotiana*, as in all plants, the two primary means by which a plant protects itself against disease are though a pre-formed, standing defense or through infection-induced responses. For each of these categories, tobacco utilizes a multitude of mechanisms, including ones common to most plant species, such as the basic cell wall structure as a pre-formed defense, the infection-induced hyper sensitive response, and pathogen-specific gene silencing mediated by native RNAi. Other disease protective schemes are more specific to tobacco, like the trichome exudates, which are leaf surface compounds that have been associated with improved flea beetle and hornworm resistance (Nielsen et al., 1982). Nicotine, the major chemical compound for which tobacco is known, leads to reduced herbivory from beet armyworms, grasshoppers, and other pests (Steppuhn et al., 2004). Brassinosteroids, a plant hormone, have been shown to activate signal cascades that play a role in defense against *Tobacco Mosaic Virus* (TMV) (Deng et al., 2016).

In general, disease resistance, especially infection-induced responses, are usually specific to a limited number of pathogen species or strains (Kim et al., 2016).

Effectors, which are compounds, small molecules, or proteins secreted by many pathogens, can modify host defenses in a way beneficial to infection spread (Rafiqi et al., 2012). To counter this, the tobacco genome possesses many resistance genes (R genes) that lead to a plant resistance response when they recognize or interact with the effector. Measuring these types of interactions can often be done using protein-protein interaction assays (ex: co-immunoprecipitation, yeast two-hybrid, or biomolecular fluorescence complementation) or indirectly though RNA-seq and differential gene expression analysis. R genes are a large diverse group, but often a single R gene will only interact with a specific effector. Due to the constantly evolving, yet similar nature of avirulence interactions, some R gene families are highly similar, leading to difficulties in assembling these families or accurately measuring the expression of a specific family member.

Due to resistance breaking strains and pathogens for which native resistance does not currently exist in commercial varieties, many methods are employed by tobacco scientists, farmers, and breeders to combat tobacco pathogens. One of the most common strategies involves the introduction/removal of specific genes or quantitative trait loci (QTL) through some form of traditional breeding or transgene-based approaches. One of the more successful means of introducing disease resistance genes into tobacco has been through the introgression of chromosomal fragments from wild *Nicotiana* species. While cross-species introgressions can transfer resistance traits not found in native *N. tabacum,*

this transfer is often accompanied by linkage drag (Wernsman, 1999). This drag often leads to yield loss or phenotypic changes detrimental to the commercial value of the plant (Wernsman, 1999). Additionally, while some *Nicotiana* species may possess beneficial resistance genes, crossing them with *N. tabacum* to transfer the trait is often not possible due to genetic distance between the species.

Transgenic strategies to introduce disease resistance in tobacco has proven to be very effective (Wernsman, 1999); however, due to the controversies surrounding genetically modified (GM) organisms, there are currently no disease resistant GM tobacco varieties in commercial production. Classically, the most common method for transgenic tobacco production has been through the *Agrobacterium tumefaciens*-based binary vector system. The binary vector system works with one vector carrying the genes required for transferal of the T-DNA into the host cell, while the second vector contains the foreign DNA to be introduced into the plant (An et al., 1986). Emerging genome editing technologies hold promise for gaining greater acceptability than standard GM approaches, as they can be used in a manner whereby no foreign DNA remains in the genome once the desired editing has taken place. While the list of viable genome editing tools include meganucleases, zinc-finger nucleases, and transcription activator-like effector nucleases (TALENs) the fastest growing method of genome modification in plants is mediated by the CRISPR/Cas9 system (Gao et al., 2015). Briefly, this method utilizes two components to edit the plant genome, an enzyme to cut the host DNA at the desired location (Cas9), and an RNA guidance system (gRNA) to target Cas9 to the specific genome location. These are the only tools needed to

introduce targeted mutations in the form of small insertions/deletions into a gene or genes of interest through utilization of the nonhomologous end joining repair machinery of the plant cell. A third component, a template from which repair is to be directed, is also required for more complex editing applications requiring the use of the homologous recombination-mediated repair. Using a template designed by a researcher to match a portion of the target plant DNA, areas of plant DNA can be edited very precisely, or foreign DNAs inserted in a predetermined location if desired.

In general, quantitative disease resistance traits result in a lower level of resistance to pathogens than can be attained via a single strong R gene, but are less easily overcome by the pathogen. Characterizing and parsing a QTL that provides resistance can be very beneficial but is often full of challenges. After markers tracking a QTL have been produced, they can be used for marker assisted breeding. For further QTL characterization, recombination to narrow the chromosomal region involved, and expression analysis, such as RNA-seq may be used to refine and identify a set of candidate genes associated with the QTL. Nevertheless, the precise characterization of the specific gene(s) involved can still be problematic, often because the nature of these genes can vary greatly, unlike strong single-gene resistance which is typically mediated by a member of the R gene family. Adding to gene identification difficulties, QTLs can be made up of genes that provide only partial protection or only function in the presence of an unlinked gene product. Additionally, if the gene is only transcribed in response to elicitation by the pathogen, expression analysis can become more complicated. Even if a classical R-gene is present in the QTL, R-gene families

are often highly homologous and therefore expression differences or differential SNPs can be masked when conducting expression analyses. If a draft genome is available for the species, the high sequence identity shared among R-genes can lead them to be under or incorrectly represented in the genome assembly (Baker, 2012).

Potyviruses

Potyviruses, of the family Potyviridae are destructive pathogens for many economically important crop species. The potyvirus, potato virus Y (PVY) can cause large losses in infected fields of tobacco, tomato, pepper, and potato. A number of other potyviruses, such as tobacco etch virus (TEV) and tobacco vein mottling virus (TVMV) can also infect tobacco. PVY is commonly spread through an aphid vector, with the green peach aphid (*Myzus persicae*) being most effective at transfer (Warren et al., 2005). However, PVY does not replicate within the aphid, and can be cleared after a number of feedings (Warren et al., 2005). PVY is comprised of three main classes of strains: PVY^{o}, PCY^{c} (some of which are not aphid transmitted), and PVY^{N} (Hussain et al., 2016). Depending on viral strain, symptoms include leaf mottling, yellowing, and veinal necrosis (in the case of PVY^{N}). Potyviruses are positive sense ssRNA viruses generally just under 9.8kb in size. A single extended open reading frame is flanked by a 3'-poly-A tail and a non-translated 5'-terminal region (5'-NTR) which is covalently linked to a viral genome-linked protein (VPg) (Figure 1). This open reading frame produces a polyprotein which is proteolytically cleaved to produce 10 of the 11 proteins encoded by the virus. The 11^{th} protein results from a frame-shifted

open reading frame near the middle of the strand. Assembled as a virion, it is a filamentous structure coated by ~2000 copies of the coat protein with a length of 680 - 900 nm (11 -15 nm wide) (Hussain et al., 2016).

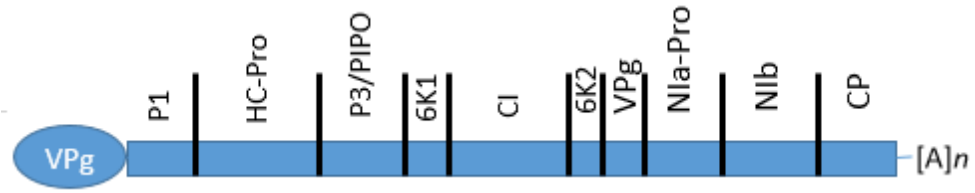


Figure 1. Potyvirus genome structure.

Structure of the potyvirus genome with covalently attached VPg (viral protein genome-linked). Proteins in order are P1, HC-Pro (helper component proteinase), P3/PIPO, 6K1, CI (cylindrical inclusion protein), 6K2, VPg, NIa-Pro (nuclear inclusion a), NIb (nuclear inclusion b), and CP (coat protein).

Steps in potyvirus infection

Potyvirus infections begin after the introduction of the viral capsid into the host cytoplasm. After uncoating, host ribosomes are utilized for translation into a polyprotein. During translation, the VPg binds to a number of host proteins, including eIF(iso)4E, PABP, eEF1A, RH8, and eIF4E (Ivanov et al., 2014). During normal host translation, eukaryotic translation initiation factor 4E (eIF4E) (and the closely related eIF(iso)4E) helps direct ribosomes to the mRNA cap structure and is believed to assist in translational efficiency and mRNA stabilization. Interestingly, VPg has been shown to not be required for potyvirus translation *in vitro*, as internal ribosome entry site (IRES)-mediated translation occurs using

the viral 5'UTR as a ribosome binding site (Gallie, 2001). While VPg may not be necessary for replication, interaction between eIF4E and VPg has been shown to be required for a successful viral infection, likely due to eIF4E's translational efficiency and mRNA stabilization functions (Robaglia and Caranta, 2006). The necessity of the eIF4E and VPg interaction to facilitate a successful viral infection is evidenced by several examples in plants where the mutation or complete loss of a host eIF4E gene has led to potyvirus resistance (Julio et al., 2015; Piron et al., 2010). The virus may also induce increased transcription of genes encoding specific ribosomal subunits, as shown by selective expression changes in infected *N. benthamiana* (Dardick, 2007). Host acidic ribosomal protein P0 is upregulated in infected cells and leads to greater virus accumulation, yet is not associated with viral spread (Hafren et al., 2013).

Potyvirus replication occurs within viral replication complexes (VRCs), whose formation is initiated by the viral 6K2-VPg-Pro precursor (Ivanov et al., 2014). Viral proteins CI, 6K2, NIa, Nib, HC-Pro, and P3 all play a role in replication, including the co-opting of host proteins and attachment of the VRC to the chloroplast (host SNARE protein Syp71 assists in chloroplast attachment (Wei et al., 2013)). Negative-strand synthesis initiates first, followed by the regeneration of the positive-strand; both processes occur within the micro-environment of the VRC. Packing and release of progeny RNA subsequently occurs. Potyvirus proteins CI, CP and P3N-PIPO are known to be involved in cell to cell movement but it is not yet known how the movement occurs, though some evidence suggests the virus is transported within viral produced vesicles through plasmodesmata (Ivanov et al., 2014).

Additionally, the viral genome must be repacked in coat protein to survive in the extracellular environment.

Potyvirus resistance mechanisms in tobacco

Within tobacco production fields, often the only defense strategy utilized against PVY is management of the local aphid population (generally though the use of pesticides and mineral oils), and reducing contact and proximity with infected plants. Through breeding, a number of genes and QTLs that improve resistance to potyviruses have been found and bred into tobacco cultivars. Resistance to various potyviruses often involve interactions with specific cellular proteins, so currently known resistance mechanisms do not protect against all potyvirus or even all strains within a given potyvirus type, such as PVY (Lewis, 2005). Through the study of a potyvirus resistance locus referred to as *va,* it has been shown that the absence or mutation of a specific isoform of eIF4E (eIF4E.1S) provides a high level of resistance to potyviruses including PVY, TEV, and TVMV (Julio et al., 2015). The *va* locus is a large deletion mutation that was presumably derived from an even larger deletion termed *VAM* (Koelle, 1961; Miller, 1987). Both loci are associated with potyvirus resistance, as elaborated further in Chapter 2 of this thesis. In addition to the *VAM* locus *per se*, the PVY resistance associated with the original line in which *VAM* was discovered carries another, unlinked mutation denoted *pvy2,* which may restrict virus accumulation (Acosta-Leal and Xiong, 2008). Another potentially unique form of PVY resistance from *N. africana* has been introduced into *N. tabacum* (Lewis, 2005) and a tobacco gametoclonal variant

designated NC602 has been shown to display some degree of PVY resistance (Witherspoon et al., 1991). Finally, in a survey of 163 known PVY resistant accessions from a large tobacco germplasm collection, Julio et al. (2015) showed that 13 of these contained wild type eIF4E.1S genes, meaning additional uncharacterized sources of resistance are likely present that could be exploited.

Understanding which tobacco genes have altered expression patterns after potyvirus infection can help during genome-based investigative studies designed to understand the molecular basis of the susceptibility/resistance responses to these viruses. In tobacco, PVY infection has been found to alter the expression of 8,133 mRNAs (Guo et al., 2017). In general, resistant lines showed upregulation of pathogenesis-related (PR) genes, and genes associated with biotic stress response, cell rescue, and defense compound synthesis. In contrast, susceptible tobaccos showed upregulation of genes associated with cell signaling, and those encoding WRKY transcription factors (Baebler et al., 2009). HSP70, a cellular chaperone involved in proper protein folding is also upregulated and may help with the processing of the increased viral protein load (Aranda et al., 1996). Although it is unknown whether potyviruses can influence host gene expression directly, viral proteins NIa, NIb, HC-Pro, and P3 have been found within host nuclei, despite that fact that virus replication occurs in the cytoplasm. Host plants exposed to abiotic stress can have reduced resistance to certain potyvirus infection and accumulation (Prasch and Sonnewald, 2013). Some evidence also points to potyvirus-mediated upregulation of cellular abiotic stress-related genes, as they may be beneficial to enhancing the infection response (Ivanov et al.,

2014). Therefore, breeding plants with higher abiotic stress resistance may indirectly increase viral resistance. Overall, changes at the protein level are not as pronounced as at the transcription level in susceptible plants (Ivanov et al., 2014).

Blue mold

Blue mold, or *Peronospora tabacina* D.B. Adam was first described 1863 (*Peronospora hyoscyami* de Bary). It is an obligate biotrophic pathogen with the ability to cause severe damage in tobacco fields. *P. tabacina* is an oomycete which, while similar to other fungal pathogens, is more closely related to brown algae and diatoms. The pathogen is a downy mildew and is present in warmer growth climates, though its spores can travel far distances on air currents to more northerly fields (LaMondia and Aylor, 2001; Blanco et al., 2017). As symptoms may take several days to become apparent, the pathogen is often spread through distribution of infected plants. Infection is generally identified by a blueish downy mold with light brown necrotic patches on the undersides of leaves (Sukanya and Spring, 2013). An infection can spread quickly through a field once spores are produced, with individual plants developing a systemic infection often leading to plant death or weakening upon *P. tabacina* reaching the vascular system.

While most downy mildews have a narrow host range, *P. tabacina* infection requires very specific conditions. Infections can be infrequent and highly dependent on the weather (Spring et al., 2013). The pathogen cannot overwinter in northern fields and must be re-introduced from wind-blown spores (Spurr and Todd, 1982). While the *P. tabacina* spores

can survive a wide range of environmental conditions, germination is most successful during times of high humidity, mean daily temperatures of 20°C, and low sunlight/UV (Sukanya and Spring, 2013).

Tobacco at any developmental stage can be infected with *Peronospora* when a spore contacts a leaf (Sukanya and Spring, 2013). Successful spore germination involves introducing germ tubes into the leaf though stomata, or utilizing appressoria. Nutrients for growth are gathered when mesophyll, bundle sheath, and epidermal cells are penetrated by filiform haustoria. Spread within the leaf occurs through hyphae propagation where an infected vascular system can lead to further translocation of the disease. Finally, sporulation generally occurs on the abaxial leaf surface when sprangiophores emerge through the stomata (Sukanya and Spring, 2013). Through oospores, sexual reproduction can occur but most *P. tabacina* spread is due to asexual or clonal propagation (sporangiospores and conidiospores).

Currently, the genomic resources for blue mold include a microsatellite investigation of the pathogen (Sukanya and Spring, 2013) and genome sequences (including the mitochondrial genomes) of *P. tabacina* strains 968-J2 and 968-S26 (Derevnina et al., 2015).

Resistance to Blue Mold in tobacco

Natural resistance against *P. tabacina* infection in commercial tobacco varieties is poor. However, for many tobacco cultivars and *Nicotiana* species that display some level of resistance, the first line of defense against *P. tabacina* infection is through leaf surface

compounds (Shepherd et al., 2005). Due to secreting trichomes on tobacco leaf surfaces, there is an abundance of compounds that have the potential to interact with the oomycete. In large enough concentrations (seen naturally in some tobacco cultivars), the secreted cembratriene diols (α-CBT-diols and β-CBT-diols), labdenediol, and sclareol have all been reported to inhibit *P. tabacina* spore germination (reviewed in Kennedy et al., 1992). In addition, proteins produced and secreted from trichomes called T-phylloplanins inhibit *P. tabacina* infection (Shepherd et al., 2005). When the gene T-*phylloplanin,* which produces one of these proteins, is expressed in the apoplast of transgenic plants, enhanced resistance is observed (Kroumova et al., 2013).

Traditionally, blue mold management has been in the form of one of the few effective externally applied chemical treatments. However, the widely used fungicides mexalaxyl and dimethomorph are showing less effectiveness than in the past (Derevnina et al., 2015**;** LaMondia, 2013]. New fungicides that use β-aminobutyric acid (Piekna-Grochala and Kepczynska, 2013) or nanoparticles comprised of Zn and ZnO (Wagner et al., 2016) may be good complements, but the discovery and implementation of new genetic sources of resistance in addition to fungicides is a better option. One such source of resistance is found within an introgression fragment that was bred into *N. tabacum* from the wild species *N. debneyi* (Clayton, 1967; Milla et al., 2005). The protection conferred by this introgression fragment (which works best when in the homozygous state) is less than that observed in the original wild species but is very useful in areas where blue mold is a common occurrence (Milla et al., 2005). Additionally, some tobacco genes have been shown to effect

development of the disease. Among these are genes encoding β-Ionone, PR-1a, glucanase, gluthione synthetase, an EIL2 transcription factor, and the above-mentioned T-phylloplanin (reviewed in Borras-Hidalgo et al., 2009).

Many of the mechanistic details of *P. tabacina* infection are unknown, but oomycetes in general have evolved tactics to breakdown plant defenses and counter resistance. Secreted PR proteins, known as effectors, influence extracellular targets in the plant apoplast or within cellular cytoplasm (Kamoun, 2006). The best characterized cytoplasmic effectors are those containing RxLR or LxLFAK motifs (Morgan et al., 2007). Cell surface effectors include small cysteine-rich proteins, elicitins, proteases and ethylene inducing proteins (NEPs) (Derevnina et al., 2015). *N. tabacum* responses to *P. tabacina* in the field are dependent on many, often uncontrollable factors such as humidity levels, cloud cover, and the physiological status of the plant. While the disease occurs yearly in the Caribbean region, migration of the pathogen into the eastern US is dependent of environmental and human factors. These factors can make finding, selecting for, and studying genetic sources of blue mold resistance in the field difficult.

Leaf trichomes and exudates

Many compounds produced within a tobacco leaf are exuded and accumulate upon the surface. A majority of the compounds transported to the tobacco leaf surface originate within trichomes, though some may leach passively from the interior of the leaf (Wagner et al., 2004). The layer of chemicals present on the leaf surface are important for protecting

against pathogens, moderating leaf temperature, protecting against UV exposure, as well as having potential commercial applications (Sallets et al., 2014). The two types of glandular trichomes found on *N. tabacum* leaves are short glandular trichomes (SGTs) and tall glandular trichomes (TGTs). It is likely that a majority of chemical secretions originate in the TGTs though details on the differing function of each are not fully understood (Sallets et al., 2014). TGTs exist as multi-celled stalks upon which a head of one or more cells rests. Additionally, unlike the SGTs, which contain a single stalk and head cell, the TGT head cells also contain chlorophyll, and thus functional chloroplasts (Sallets et al., 2014). Some of the compounds exuded through the TGTs are produced within the head cells of the trichome, while others, such as nicotine, can be transported to TGTs for secretion (Wagner et al., 2004). Upon secretion, many on the compounds are retained between the head cells and a waxy cuticle layer surrounding the head (Kandra and Wagner, 1988).

N. tabacum trichomes exude many, mostly hydrophobic compounds including fatty acids, alkanes, sugar esters, and terpenoids (Sallets et al., 2014). A number of studies, including those measuring the transcriptomes and proteomes of isolated trichomes, have shown that trichomes play a large role in the production of secondary metabolites destined for exudation (Cui et al., 2011; Sallets et al., 2014). A lipid transport protein and ABC transporters have been found to assist in diterpene mobilization in *N. tabacum* trichomes (Yong Eui et al., 2012; Jerome et al., 2013). The diterpenes mostly commonly found on the leaf surface of tobacco are the cembranoids, α- and β-CBT diols (referred to as α- and β-DVT

diols in earlier literature), Z-abienol (sometimes also referred to as *cis*-abienol) and labdene-diol (Jassbi et al., 2017).

The production of *N. tabacum* diterpenes begins with geranylgeranyl diphosphate (GGPP), synthesized from dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP) (Yan et al., 2016). The four major end products of this pathway, which starts with GGPP, are α- and β-CBT diols, Z-abienol, and labdene-diol, as shown in Figure 2 (Sallaud et al., 2012). GGPP is converted to α- and β-CBT-ols through a cemrantrien-ol synthase (CBTS) catalyzed reaction, followed by a cytochrome P450 hydroxylase (CYP71D16) reaction to create α- and β-CBT-diols (Wang et al., 2001). Z-abienol and labdene-diol production are catalyzed by Z-abienol cyclase and labdene-diol cyclase respectively, using an 8- α-hydroxy-copalyl-pyrophosphate (8-OH-CPP) intermediate (Yan et al., 2016; Sallaud et al., 2012). The production of these diterpenes appears to be light dependent (Kandra and Wagner, 1998) requiring functional chloroplasts within TST heads (Nielsen et al., 1982; Nielsen and Severson, 1990).

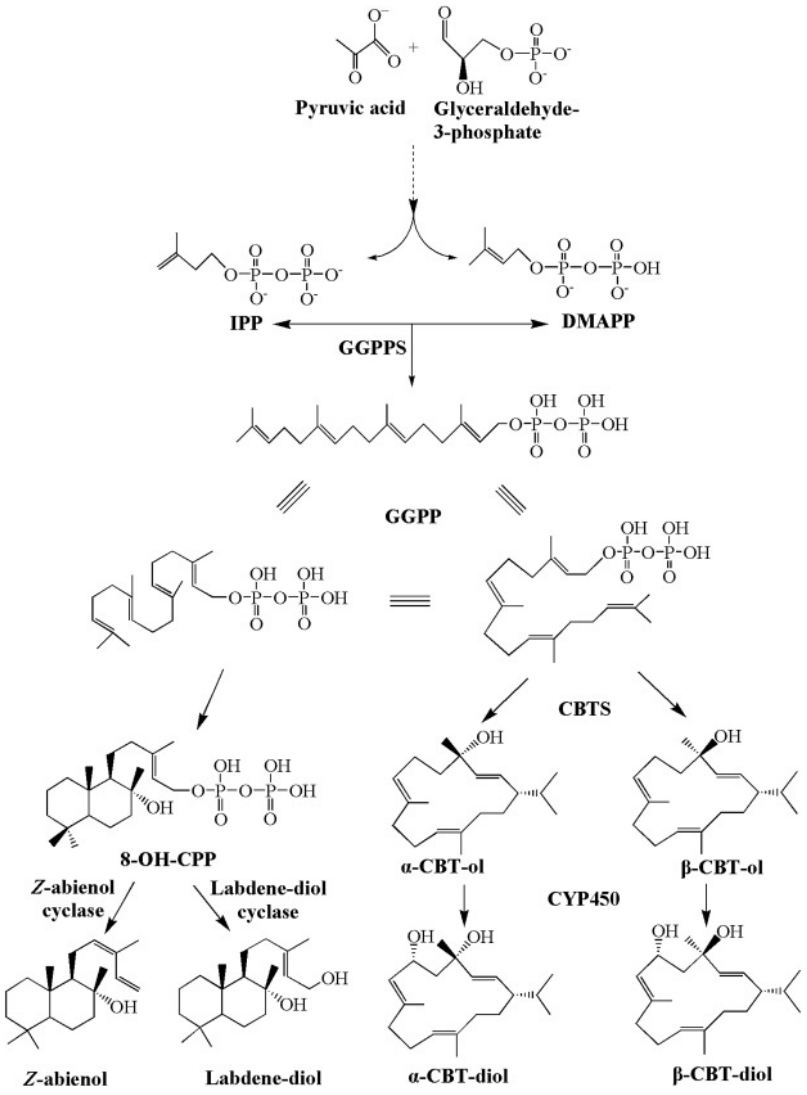


Figure 2: Major tobacco diterpenoid biosynthetic pathway (Image from Yan et al., 2016).

This thesis describes the use of NGS technologies, combined with the most current tobacco genome resources to investigate the structure and gene composition of three important genetic loci of significant historical and commercial value. Chapter 2 describes the use of RNA-seq and publically available genome sequence resources to characterize the

VAM deletion mutant and the related shorter deletion mutation *va*, including a study of the genes within the chromosome deletion associated with potyvirus susceptibility and the synthesis of leaf surface cembranoids. Chapter three involves the characterization of an introgression fragment derived from *N. debneyi* that contains a strong blue mold resistance QTL, and efforts to shorten the size of the introgression fragment (to alleviate yield drag) while retaining the disease resistance QTL.

REFERENCES

Acosta-Leal R, Xiong Z. Complementary functions of two recessive R-genes determine resistance durability of tobacco 'Virgin A Mutant' (VAM) to Potato virus Y. Virology. 2008;379(2):275–283.

An G, Watson BD, Chiang CC. Transformation of tobacco, tomato, potato, and *Arabidopsis thaliana* using a binary Ti vector system. Plant Physiology. 1986;81:301-305.

Aranda MA, Escaler M, Wang D, Maule AJ. Induction of HSP70 and polyubiquitin expression associated with plant virus replication. PNAS. 1996;93:15289-15293.

Baebler S, *et al.* PVY^{NTN} elicits a diverse gene expression response in different potato genotypes in the first 12 h after inoculation. Mol Plant Pathol. 2009;10:263-275.

Baker M. *De novo* genome assembly: what every biologist should know. Nat Methods. 2012;9:333-337.

Bendahmane A. An induced mutation in tomato eIF4E leads to immunity to two potyviruses. PloS ONE. 2010;5(6).

Bennett MD, Leitch IJ. Nuclear DNA amounts in angiosperms. Ann Bot. 1995;76:113- 76.

Bindler G, Plieske J, Bakaher N, Gunduz I, Ivanov N, Van der Hoeven R, Ganal M, Donini P. A high density genetic map of tobacco (Nicotiana tabacum L.) obtained from large scale microsatellite marker development. Theor Appl Genet. 2011;123:219–230.

Blanco M, Carbone I, Ristaino J. Population structure and migration of the Tobacco Blue Mold Pathogen, *Peronospora tabacina* into North America and Europe. Mol. Ecol. 2017;In press.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–2120.

Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller A, Martin GB. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Molecular Plant-Microbe interactions. 2012;25:1523-1530.

Borras-Hidalgo O, Thomma BPHJ, Silva Y, Chacon O, Pujol M. Tobacco blue mould disease caused by *Peronospora hyoscyami* f. sp. *tabacina*. Mol Plant Path. 2009;11:13-18.

Cai G, Liang S, Zheng X, Xiao F. Local sequence and sequencing depth dependent accuracy of RNA-seq reads. BMC Bioinformatics. 2017;18:364.

Choi YE, Lim S, Kim HJ, Han JY, Lee MH, Yang Y, Kim JA, Kim YS. tobacco NtLTP1, a glandular-specific lipid transfer protein, is required for lipid secretion from glandular trichomes. Plant J. 2012;70:480-491.

Clayton EE. The transfer of blue mold resistance from *Nicotiana debneyi*. Part III. Development of a blue mold resistant cigar wrapper variety. Tob Sci. 1967;11:107-110.

Crouzet J, Roland J, Peeters E, Trombik T, Ducos E, Nader J, Boutry M. NtPDR1, a plasma membrane ABC transporter from *Nicotiana tabacum*, is involved in diterpene transport. Plant Mol Biol. 2013;82:181-192.

Cui H, Song-Tao Z, Hui-Juan Y, Hao J, Xiu-Jie W. Gene expression profile analysis of tobacco leaf trichomes. BMC Plant Biology. 2011;11.

Dardick C. Comparative expression profiling of *Nicotiana benthamiana* leaves systemically infected with three fruit tree viruses. Mol Plant Microbe Interact. 2007;20:1007-1017.

Deng XG, Tong Z, Peng XJ, Xi DH, Guo H, Yin Y, Zhang DW, Lin HH. Role of brassinosteroid signaling in modulating *Tobacco mosaic virus* resistance in *Nicotiana benthamiana*. Scientific Rep 2016;6:20579.

Derevnina L, Chin-Wo-Reyes S, Martin F, Wood K, Froenicke L, Spring O, Michelmore R. Genome sequence and architecture of the tobacco downy mildew pathogen *Peronospora tabacina*. Mol Plant-Microbe Interactions. 2015;11:1198-1215.

Edwards KD, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans AD, Bombarely A, Allen F, Hurst R, White B, Kernodle SP, Bromley JR, Sanchez-Tamburrino JP, Lewis RS, Mueller LA. A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. BMC Genomics. 2017;18:448

Eid J. *et al*. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133-138.

Fricano A, Bakaher N, Corvo MD, Piffanelli P, Donini P, et al. Molecular diversity, population structure, and linkage disequilibrium in a worldwide collection of tobacco (Nicotiana tabacum L.) germplasm. BMC Genet. 2012;13:18.10.1186/1471-2156-13-18.

Gallie DR. Cap-independent translation conferred by the 5' leader of tobacco etch virus is eukaryotic initiation factor 4G dependent. J Virol. 2001;75:12141-12152. Gao J, Wang G, Ma S, Xie X, Wu X, Zhang X, Wu Y, Zhao P, Xia Q. CRISPR/Cas9 mediated targeted mutagenesis in *Nicotiana tabacum*. Plant Molecular Biology. 2015;87:99-110.

Gerstel DU, Burns JA, Burk LG. Interspecific hybridizations with an African tobacco, *Nicotiana africana* Merxm. Journal of Heredity. 1979;5:342-344.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Gevet. 2016;17:333-51.

Guo J, Xu N, Li Z, Zhang S, Wu J, Hyun Kim D, Marma MS, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachikov S, Russo JJ, Turro NJ, Ju J. Four-color DNA sequencing with 3'-*O*-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci USA. 2008;105:9145-9150.

Guo Y *et al*. Integrated analysis of tobacco miRNA and mRNA expression profiles under PVY infection provides insight into tobacco-PVY interactions. Sci Rep. 2017;7:4895.

Gur A, Zamir D. Unused natural variation can lift yield barriers in plant breeding. PLOS Biol. 2004;2:1610–1615.

Hafren A, Eskelin K, Makinen K. Ribosomal protein P0 promotes Potato virus A infection and funtions in viral translation together with VPg and eIF(iso)4E. J Viro. 2013;87:4302-4312.

Hussain A, Arif M, Abbas A, Hussain B, Ali M, Jaffar S. A review on aphid-borne virus (Potato Virus Y). Journal of Entomology and Zoology Studies. 2016;3:189-192.

Ivanov KI, Eskelin K, Lohmus A, Makinen K. Molecular and cellular mechanisms underlying Potyvirus infection. J Gen Virol. 2014;95:1415-29.

Jassbi A, Zare S, Asadollahi M, Schuman M. Ecological roles and biological activites of specialized metabolites from the genus *Nicotiana*. Chemical Reviews. 2017;preprint.

Julio E, Cotucheau J, Decorps C, Volpatti R, Sentenac C, Candresse T, de Borne FD. A eukaryotic translation initiation factor 4E (eIF4E) is responsible for the “va” tobacco recessive resistance to potyviruses. Plant Mol Biol Rep. 2015;33:609–623.

Kamoun S. A catalogue of the effector secretome of plant pathogenic oomycetes. Annual review of phytopathology. 2006;44:41-60.

Morgan W, Kamoun S. RXLR effectors of plant pathogenic oomycetes. Current Opionion in Microbiology. 2007;4:332-338.

Kandra L, Wagner GJ. Studies of the site and mode of biosynthesis of tobacco trichome exudate components. Arch Biochem Biophys. 1988;265:425-432.

Kennedy BS, Nielsen MT, Severson RF, Sisson VA, Stephenson MK, Jackson DM. Leaf surface chemicals from *Nicotiana* affecting germination of *Peronospora tabacina* (adam) sporangia. Journal of Chemical Ecology. 1992;9:1467-1479.

Kenton A, Parokonny AS, Gleba YY, Bennett MD. Characterization of the *Nicotiana tabacum* L. genome by molecular cytogenetics. Mol Gen Genet. 1993;240:159-69.

Kim SH, Qi D, Ashfield T, Helm M, Innes RW. Using decoys to expand the recognition specificity of a plant disease resistance protein. Science. 2016;6274:684-687.

Koelle G. Genetic analyse einer Y-virus (Rippenbraune) resistenten mutante der tabaksorte Virgin A. Zuchter. 1961;31:71–71.

Kroumova A, Sahoo D, Raha S, Goodin M, Maiti I, Wagner G. Expression of an apoplast directed, T-*phylloplanin*-GFP fusion gene confers resistance against *Peronospora tabacina* disease in a susceptible tobacco. Plant Cell Reports. 2013;11:1771-1782.

LaMondia JA. Reduced sensitivity of *Peronospora tabacina*, causal agent of tobacco blue mold, to Dimethomorph fungicide in Connecticut. Tob Sci. 2013;50:19-24.

LaMondia JA, Aylor DE. Epidemiology and management of a periodically introduced pathogen. Biological Invasions. 2001;3:273-282.

Lewis RS, Nicholson JS. Aspects of the evolution of *Nicotiana tabacum* L. and the status of the United States *Nicotiana* Germplasm Collection. Gen Res and Crop Evo. 2007;54:727-740.

Lewis RS. Transfer of resistance to potato virus Y (PVY) from Nicotiana africana to Nicotiana tabacum: possible influence of tissue culture on the rate of introgression. Theor Appl Genet. 2005;110:678-687.

Mascher M, Stein N. Genetic anchoring of whole-genome shotgun assemblies. Front Genet. 2014;5:208.

Mila A, Radcliff J. Flue-cured tobacco disease report. http://www.ncsu.edu/project/tobaccoportal/wp-content/uploads/2010/12/2010-Disease-Reports.pdf.

Milla SR, Levin JS, Lewis RS, Rufty RC. RAPD and SCAR markers linked to an introgressed gene conditioning resistance to *Peronospora tabacina* D.B. Adam. in tobacco. Crop Sci. 2005;45:2346-2354.

Miller RD. Registration of TN 86 burley tobacco. Crop Sci. 1987;25:2,365-366.

Murphy JP, Cox TS, Rufty RC, Rodgers DM. A representation of the pedigree relationships among flue-cured tobacco cultivars. Tob Sci. 1987;31:70-75.

Nielsen MT, Akers CP, Jarlford VE, Wagner GJ, Berger S. Comparative ultrastructural features of secreting and non-secreting glandular trichomes of two genotypes of N. tabacum. Bot Gaz. 1991;152:13-22.

Nielsen MT, Jones GA, Collins GB. Inheritance pattern for secreting and nonsecreting glandular trichomes in tobacco. Crop Sci. 1982;22:1051–1053.

Nielsen MT, Severson RF. Variation for flavor components on leaf surfaces of tobacco genotypes differing in trichome density. J Agric. Food Chem. 1990;38:467-471.

Piekna-Grochala J, Kepczynska E. Induction of resistance against pathogens by β aminobutyric acid. Acta Physiologiae Plantarum. 2013;6:1735-1748.

Piron F, Maryse N, Minoïa S, Piednoir E, Moretti A, Salgues A, Zamir D, Caranta C,

Prasch CM, Sonnewald U. Simultaneous application of heat, drought, and virus to *Arabidopsis* plants reveals significant shifts in signaling networks. Plant Physiol. 2013;162:1849-1866.

Rafiqi M, Ellis JG, Ludowici VA, Hardham AR, Dodds PN. Challenges and progress towards understanding the role of effectors in plant-fungal interactions. Curr Opin Plant Biol. 2012;15:477-482.

Robaglia C, Caranta C. Translation initiation factors: a weak link in plant RNA virus infection. Trends Plant Sci. 2006;11:40-45.

Sallaud C, Giacalone C, Töpfer R, Goepfert S, Bakaher N, Rösti S, Tissier A. Characterization of two genes for the biosynthesis of the labdane diterpene Z-abienol in tobacco (Nicotiana tabacum) glandular trichomes. Plant J. 2012;72:1–17.

Sallets A, Beyaert M, Boutry M, Champange A. Comparative proteomics of short and tall glandular trichomes of *Nicotiana tabacum* reveals differential metabolic activities. J Proteome Res. 2014;7:3386-3396.

Shepherd R, Bass TW, Houtz RL, Wagner GJ. Phylloplanins of tobacco are defensive proteins deployed on aerial surfaces by short glandular trichomes. Plant Cell. 2005;17:1851-1561.

Sierro N, Battey JND, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV. The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun. 2014;5:3833.

Sierro N, et al. Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. Genome Biol. 2013;14:R60.

Sierro N, Oeveren JV, van Eijk MJT, Martin F, Stormo KE, Peitsch MC, Ivanov NV. Whole genome profiling physical map and ancestral annotation of tobacco Hicks Broadleaf. Plant J. 2013b;75:880-889.

Spring O, Hammer TR, Zipper R, Billenkamp N. Population dynamics in tobacco blue mold incidences as a consequence of pathogen control and virulence performance of *Peronospora tabacina* phenotypes. Crop Prot. 2013;45:76-82.

Spurr HW, Todd FA. Oospores in blue mold diseased North Carolina Burley and Flue-cured tobacco. Tobacco Science. 1982;26:44-46.

Steppuhn A, Gase K, Krock B, Halitschke R, Baldwin IT. Nicotine's defensive function in nature. Plos Biol. 2004;2:e217.

Sukanya SL, Spring O. Influence of temperature and ultra-violet light on viability and infectivity of *Peronospora tabacina* sporangia. Crop Protection. 2013;51:14-18.

Trigiano RN, Wadl PA, Dean D, Hadziabdic D, Scheffler BE, Runge F, Telle S, Thines M, Ristaino J, Spring O. Ten polymorphic microsatellite loci identified from a small insert genomic library for *Peronospora tabacina*. Mycologia. 2012;104:633-640.

Wagner G, Korenkov V, Judy JD, Bertsch PM. Nanoparticles composed of Zn and ZnO inhibit *Peronospora tabacina* spore germination *in vitro* and *P. tabacina* infectivity on tobacco leaves. Nanomaterials. 2016;3:50.

Wagner GJ, Wang E, Shepherd RW. New approaches for studying and exploiting an old protuberance, the plant trichome. Annals of botany. 2004;93:3-11.

Wang E, Wang R, Deparasis J, Loughrin JH, Gan S, Wagner GJ. Suppression of a P450 hydroxylase gene in plant trichome glands enhances natural-product-based aphid resistance. Nature Biotechnology. 2001;19:371–374.

Warren M, Kruger K, Schoeman AS. Potato virus Y (PVY) and Potato Leafroll virus (PLRV). A South African perspective. University of Pretoria. 2005;32pp.

Wei T, Zhang C, Hou X, Sanfacon H, Wang A. The SNARE protein Syp71 is essential for turnip mosaic virus infection by mediating fusion of virus-induced vesicles with chloroplasts. PLoS Pathog. 2013;9:e1003378.

Wernsman EA. An overview of tobacco breeding - past, present and future. Rec Adv Tob. 1999;25:5-35.

Witherspoon WD, Wernsman EA, Gooding GV Jr, Rufty RC. Characterization of a gametoclonal variant controlling virus resistance in tobacco. Theor Appl Genet. 1991;81:1-5.

Wu X, Li D, Bao Y, Zaitlin D, Miller R, Yang S. Genetic dissection of disease resistance to the blue mold pathogen, *Peronospora tabacina*, in tobacco. Agronomy. 2015;4:555-568.

Xu S, et al. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. PNAS. 2017;114:6133-6138.

Yan N, Du Y, Liu X, Zhang H, Lin Y, Zhang P, Gong D, Zhang Z. Chemical structures, biosysnthesis, bioactivities, biocatalysis and semisynthesis of tobacco cembranoids: An overview. Industrial Crops and Products. 2016;83:66-80.

CHAPTER 2

Characterization of *Nicotiana tabacum* Genotypes Possessing Deletion Mutations that Affect Potyvirus Resistance and the Production of Trichome Exudates

Kurtis L. Dluge[1], Zhongbang Song[2], Bingwu Wang[2], Yong Liu[2*] and Ralph E. Dewey[1*]

1. Department of Crop Science, North Carolina State University, Raleigh, NC 27695, USA
2. Yunnan Academy of Tobacco Agricultural Sciences, No. 33 Yuantong St. Kunming 650021, P.R. China

*Correspondance: ralph_dewey@ncsu.edu; yliu@yntsti.com

E-mail addresses of other authors:

Kurtis L. Dluge, kldluge@ncsu.edu

Zhongbang Song, zbsoon@vip.163.com

Bingwu Wang, bwwang76@hotmail.com

This chapter is formatted for submission to BMC Biology

ABSTRACT

Background

Advances in genomics technologies are making it increasingly feasible to characterize breeding lines that carry traits of agronomic interest, even in crop species with large polyploid genomes such as tobacco. Tobacco germplasm lines that carry loci designated *VAM* and *va* have been extensively investigated due their association with potyvirus resistance (both *VAM* and *va*) and defects in leaf surface compounds originating from glandular trichomes (*VAM* only). Molecular studies and classical genetic analyses are consistent with the model that *VAM* and *va* represent deletion mutations in the same chromosomal region. In this study, we used RNA-seq analysis, together with emerging tobacco reference genome sequence data to characterize the genomic regions deleted in tobacco lines containing *VAM* and *va*.

Results

Tobacco genotypes TI 1406 (*VAM*), K326-*va* and K326 (wild type) were analyzed using RNA-seq to generate a list of genes differentially expressed in TI 1406 and K326-*va*, versus the K326 control. Candidate genes were localized onto tobacco genome scaffolds and validated as being absent in only *VAM*, or missing in both *VAM* and *va,* through PCR analysis. These results enabled the construction of a map that predicted the relative extent of the *VAM* and *va* mutations on the distal end of chromosome 21. The RNA-seq analyses lead to the discovery that members of the cembratrienol synthase gene family are deleted in TI 1406, a result consistent with the observation that plants containing the *VAM* locus

lack α- and β- cembratrien-diols. Transformation of TI 1406 with a cembratrienol synthase cDNA, however, did not recover the leaf chemistry phenotype. Common to both TI 1406 and K326-*va* was the absence of a gene encoding a specific isoform of a eukaryotic translation initiation factor (eiF4E1.S). Transformation experiments showed that ectopic expression of eiF4E1.S is sufficient to restore potyvirus susceptibility in plants possessing either the *va* or *VAM* mutant loci.

Conclusions

We have demonstrated the feasibility of using RNA-seq and emerging whole genome sequence resources in tobacco to characterize the *VAM* and *va* deletion mutants. These results lead to the discovery of genes underlying some of the phenotypic traits associated with these historically important loci. Additionally, initial size estimations were made for the deleted regions, and dominant markers were developed that are very close to one of the deletion junctions that defines *va*.

BACKGROUND

For most major crop species, a vast array of useful genetic diversity is represented in germplasm collections and/or breeding lines [1]. Deciphering the unique genetic features of these lines can lead to improvements in plant quality, biotic and abiotic stress resistance, and the modification of pathways that lead to the production of novel compounds. Identification of the specific genes underlying these traits is still difficult, particularly in polypoid species such as tobacco (*N. tabacum*), which contain large, repetitive genomes and

incomplete genomic resources. Recently, parsing and characterizing unique genes of interest in tobacco has become increasingly achievable with streamlined next generation sequencing protocols and the availability of additional tobacco genomic resources [2].

Tobacco Introduction (TI) 1406 contains or Virgin A Mutant (*VAM*), a well-known source of virus resistance that was originally generated through irradiation of the Virgin A cultivar [3]. Historically, its most notable attribute has been the conferral of recessive resistance toward strains of the potyviruses PVY (potato virus Y), TEV (tobacco etch virus), and TVMV (tobacco vein mottling virus) [3, 4, 5]. Plants that possess the *VAM* mutation also carry traits that make them undesirable for cultivation, most notably certain trichome exudates are either negligible or completely missing [6]. In addition, plants containing *VAM* lack chloroplasts within the trichome head and have generally smaller or thinner leaves [7]. Deficiencies in trichome exudates have been associated with decreased resistance to the tobacco flea beetle and hornworm [6].

Glandular trichome secretions in tobacco are primarily composed of two diterpenoid groups, the macrocyclic cembranoids, α and β-cembratriene-diols (α-, β-CBT-diols), and the polycyclic labdanoids, Z-abienol and labdene-diol. Although virtually all normal tobacco plants accumulate high levels of the α- and β-CBT-diols, the Z-abienol and labdene-diol levels vary widely among germplasms lines, with these compounds typically being found in greatest amounts in tobaccos of the Oriental market type [34]. A variety of sucrose esters of short chain fatty acids can also be detected in tobacco leaf surface exudates, which like Z-abienol, are typically found in greatest abundance in Oriental-type tobaccos [14, 15].

Many of the genes encoding enzymes involved in the formation of tobacco trichome exudates have been characterized in *N. tabacum*. The production of CBT-diols is initiated with the cyclization of geranylgeranyl diphosphate (GGPP) into α- and β-CBT-ols by the enzyme cembratrienol synthase (CBTS). RNAi-mediated inhibition of a CBTS-encoding gene designated *CYC-1* was very effective in inhibiting the production of α- and β-CBT-ols, and α- and β-CBT-diols [16]. After the CBTS-mediated production of α- and β-CBT-ols, a cytochrome P450 enzyme encoded by the *CYP71D16* gene hydroxylates these compounds to form α- and β-CBT-diols [17, 18]. The CYP71D16-mediated reaction is very efficient, as only trace amounts of the α- and β-CBT-ols can typically be detected on the leaf surface. The biosynthesis of Z-abienol also initiates from GGPP, as a class II terpene synthase encoded by the *NtCPS2* gene converts GGPP to 8-hydroxyl-copalyl diphosphate. Subsequent production of Z-abienol from 8-hydroxyl-copalyl diphosphate is catalyzed by a kaurene synthase-like enzyme encoded by the *NtABS* gene [27]. From the analysis of over 100 diverse tobacco cultivars, it was shown that debilitating mutations in *NtCPS2* were highly associated with genotypes that lacked Z-abienol as a leaf surface compound [27].

Imaging of trichome heads from TI 1406 showed a lack of exudate buildup, though the cuticle that would normally surround the exudate appeared to be present. The most dramatic ultrastructural feature distinguishing T1 1406 trichomes, however, is the lack of chloroplasts within the head; in comparison, normal tobacco trichome heads possess numerous, well-developed chloroplasts [7]. Analysis of the leaf surface components of TI 1406 showed that this genotype produced very low levels of α- and β-CBT-diols, and trace

amounts of Z-abienol and the α-methylvaleric acid-containing sucrose esters (BMVSE) [14]. Doubled haploid (DH) lines that were derived from a cross between TI 1406 and burley variety Ky14 produced a range of leaf surface chemistry phenotypes for α- and β-CBT-diols, Z-abienol and BMVSE. The observation of several DHs lines producing substantial quantities of Z-abienol and BMVSE was unexpected, given that Ky14 does not produce these compounds. The authors of this study concluded that the TI 1406 genome does in fact possess the genes capable of facilitating Z-abienol and BMVSE production, but in the presence of the *VAM* locus are unable to produce more than trace amounts of these due to the trichome secretory deficiencies caused by this mutation [14], a deficiency that they attributed to the lack of functional chloroplasts within the trichome heads of TI 1406 [7, 14].

The *va* locus originated as a single plant selection displaying a high level of potyvirus resistance that was identified from a complex F_4 breeding population [8]. Because TI 1406 was represented within the lineage, it was presumed that the *VAM* locus was the source of the resistance observed in this plant. Interestingly, unlike plants possessing *VAM*, those with *va* possess normal trichome secretions and a normal trichome chloroplast phenotype. Since the *va* locus confers desirable virus resistance in the absence of unwanted leaf chemistry alterations, it has been widely deployed in commercial tobacco varieties. Virus resistance in TI 1406 appears to be somewhat stronger than plants carrying *va* as there are potentially two loci providing PVY resistance, each independently inherited as recessive loci *va* and *va2*. These have been proposed to restrict viral intercellular movement and restrict virus accumulation, respectively [10]. Recently, potyvirus resistance attributed to *va* was shown

to be due to the absence of a specific eukaryotic translation initiation factor 4E (eIF4E) isoform [11], an isoform designated eIF4E1.S by Sierro *et al*. [2]. Specific eIF4E isoforms can facilitate infection when PVY's viral genome-linked protein (VPg) mimics the 5'-cap structure of messenger RNAs, allowing it to bind to eIF4E during the process of viral replication [12]. *Va* maps to an end of linkage group 21, in a part of the chromosome that is believed to have originated from *N. tabacum*'s maternal ancestor, *N. sylvestris* [11, 13]. Although both *VAM* and *va* represent deletion mutants that confer potyvirus resistance, it is not known if the entirety of *VAM* is simply an extension of the deletion that defines *va*. It is also unknown if *VAM* corresponds solely to the segment of chromosome 21 derived from the *N. sylvestris* progenitor, as the majority of chromosome 21 originated from the paternal *N. tomentosiformis* ancestor [13].

In the present study, we investigate the chromosomal regions and genes that are deleted in tobacco plants possessing the *VAM* and *va* loci, and examine their relationship to the virus resistance and leaf surface chemistry phenotypes. RNA-seq was used to compare tobacco lines K326, K326 containing *va* (K326-*va*), and the *VAM* carrier TI 1406 in order to produce a set of gene deletions common to both *VAM* and *va*, and those unique to *VAM*. This information was used in conjunction with recently reported draft genomes of tobacco to roughly define the location and extent of the *VAM* and *va* deletion mutations. The role of the eIF4E1.S gene in conferring potyvirus susceptibility was further characterized in transformation experiments conducted in plants possessing either the *VAM* or *va* loci. Transformation experiments were also conducted with a member of the cembratrienol

synthase gene family shown to be absent in plants possessing the *VAM* locus, to test its ability to help restore trichome exudate production in these plants.

RESULTS

Sequencing and transcriptome generation

RNA-seq that was performed on pooled young leaf tissue, generated raw reads of 30,284,433 for K326, 58,004,150 for K326-*va*, and 61,640,362 for TI 1406. Trimming and quality control left reads of 28,309,873 for K326, 46,213,114 for K326-*va*, and 52,132,200 for TI 1406. The WT K326 reads were used initially to generate a transcriptome which was filtered of contigs with less than 10-fold coverage or 500bp in length, resulting in a contig N50 of 1669 for 40,080 Trinity produced genes. Due to naturally occurring polymorphisms between the K326 and TI 1406 genomes, limiting false positives during alignment was crucial to keep deletion candidates to a manageable number. Therefore, K326 RNA-seq reads were used to construct the transcriptome, and the RNA-seq reads from TI 1406 were mapped against it. Another potential alignment option would have been to align all reads to each *N. tabacum* ancestral parent (*N. tomentosiformis* and *N. sylvestris*) transcriptome. This may have resulted in fewer gene isoforms to parse but would have likely increased misaligned reads due to further evolutionary distance, and thus was not was pursued.

Candidate contig list creation through expression analysis

The deletion mutation that defines *va* is speculated as having been derived from the larger *VAM* deletion, of which TI 1406 was the source [8]. Therefore, contigs with dramatically lower expression in K326-*va* or TI 1406, as ranked by fold-change in comparison to K326, are candidates for being genes that are deleted in both *va* and *VAM*. Contigs exclusively lower in TI 1406 are candidates for being uniquely absent in plants with the *VAM* locus. Using default BWA-MEM settings, reads from K326, K326-*va* and TI 1406 were aligned to the K326 transcriptome we generated. Alignments were then normalized for post-processing sequencing depth, and contigs were ranked by expression fold differences between K326 and each of the other cultivars. An above zero level of TI 1406 and/or K326-*va* alignments was seen for many of the best candidate contigs on the list. The existence of a non-zero level of reads for a given contig did not disqualify it as a candidate, as this could merely be the result of a degree of cross mapping that would be expected in the highly repetitive tobacco genome.

To reduce false positives, 25,629,411 trimmed RNA-seq K326 reads (SRR955772) from Sierro *et al.* [2] were also aligned to our K326 transcriptome. As both sets of K326 reads come from leaf tissue grown under similar conditions, this comparison facilitated the removal of highly variable genes. Contigs with a greater than a 2-fold difference in expression between the two independent K326 sets were not considered as *va* or *VAM* candidates.

Despite the high similarity between *N. tabacum* accessions in general, false positives may arise due to inherent differences between the K326 and TI 1406 gene orthologs. K326 is a flue-cured variety while the origins of TI 1406 are not clear. An investigation of SSR markers across numerous tobacco accessions by Fricano *et al.* [19] showed that TI 1406 may be somewhat more related to burley rather than flue-cured varieties. To help control for potential genetic differences, a set of the burley variety TN90 leaf RNA-seq contigs (SRR1199203) were downloaded, processed with Trimmomatic and assembled into contigs using the same settings as our K326 reads. The top differentially expressed TI 1406 contigs from above were aligned with the constructed TN90 transcriptome using Blast-2.2.29+. Because TN90 contains the *va* locus, genes unique to *VAM* should be present in TN90 and absent in TI 1406. Therefore, a gene with low or no expression in TI 1406 that displayed high quality alignments between K326 and TN90 transcriptomes was viewed as further evidence that the gene was not scored as under-expressed due simply to sequence divergence. Furthermore, differentially expressed candidate contigs predicted from K326-*va* to K326 transcriptome alignments were removed if they had an excellent match in TN90.

A list of candidate genes likely to be missing in *VAM*, sorted by fold-change and created using all false positive controls, contained 55 contigs whose expression change was > 10-fold in K326 over TI 1406, or completely absent in TI 1406 (Table 1). When contigs were sorted by K326 versus K326-*va* fold change, 12 were found that displayed at least a 7-fold difference. Of these, all except contig 22586c0g2i1, whose corresponding TI 1406 to K326 fold change is 7.08, were also found within the *VAM* set in Table 1. This is consistent

with the hypothesis that the entire *va* deletion was derived from *VAM*. Sequences were annotated with BLASTN when BLASTX failed to return a significant result. Among the annotated contigs in Table 1 are a heat shock protein (85964c0g1i1), a *N. tabacum* Sar8.2 gene (23466c1g2i2), and isoforms of cembratrien-ol synthase (CBTS) (33989c1g1i3). CBTS plays a major role in the production of trichome exudates and multiple isoforms, or alternative splicing variants, of this gene appear on the list.

Table 1 Contigs with greater than 10-fold (K326/TI 1406) or 7-fold (K326/ K326-*va*) differential expression

Contig ID	K326 coverage	K326/TI 1406	K326/K326 va	PCR	Annotation	S or T
24507c0g1i1	27.7	na	0.65	yes	Cembratrienol synthase 3	S
33989c1g1i3	408.3	2351.50	0.18	-	Cembratrienol synthase 3	S
33989c1g1i6	323.3	1443.83	0.06	yes	Cembratrienol synthase 2a	S
33989c1g1i4	204.4	1333.67	0.11	yes	Cembratrienol synthase 2a	S
33989c1g1i5	204.9	952.75	0.14	yes	Cembratrienol synthase 2a	S
33989c1g1i2	221.1	463.05	0.10	yes	Cembratrienol synthase 2a	S
8697c0g1i1	83.2	447.40	158.00	-	Uncharacterized protein LOC104224958	S
87014c0g1i1	59.5	203.87	-0.08	-	-	
23542c0g2i2	31.1	183.15	0.20	-	-	
22222c0g1i1	296.4	127.47	29.46	-	Auxin-repressed 12.5 kDa protein-like	S
33989c1g2i1	141.4	113.38	0.36	yes	Cembratrienol synthase 2a	S
22534c0g3i1	21.2	79.41	29.55	yes*	Uncharacterized protein LOC104232912	S
35943c0g2i1	24.3	48.72	35.73	yes*	3-Isopropylmalate dehydratase	S
23466c1g2i1	656.1	47.47	-0.39	yes	Sar8.2c	T
2175c0g2i1	55.9	42.48	24.70	-	Trehalose-phosphate phosphatase G	S
22475c0g1i2	56.0	36.96	18.54	yes*	Cyclic phosphodiesterase-like	S
2175c0g1i1	56.6	33.51	22.45	-	Trehalose-phosphate phosphatase G	S
33215c0g1i1	75.0	33.11	0.18	-	Uncharacterized protein LOC107810047	
64746c0g1i1	22.7	32.65	327.11	-	-	

Table 1 continued

28889c6g4i1	34.6	30.27	0.40	-	Uncharacterized protein LOC107777218	T
42740c0g1i1	40.8	25.89	0.24	yes	Uncharacterized protein LOC107765064	T
27218c1g1i1	10.3	25.31	0.30	-	Uncharacterized protein LOC107786237	S
32651c2g7i3	10.8	25.24	-0.18	-	F-box protein	T
65564c0g1i1	29.0	23.74	0.05	-	-	
23466c1g2i2	966.4	23.70	-0.36	yes	Sar8.2k	T
25192c0g1i1	34.1	22.46	0.34	-	Bifunctional purple acid phosphatase 26	S
25436c0g1i2	19.4	22.33	0.22	yes	Uncharacterized protein LOC104238501	S
8352c0g1i1	10.9	21.95	0.79	-	Uncharacterized protein LOC107812316	T
25661c1g1i3	42.9	19.09	11.64	yes*	Cyclic phosphodiesterase	S
40439c0g1i1	10.7	18.95	0.19	-	-	
22779c0g1i1	118.1	18.95	17.29	-	heat- and acid-stable phosphoprotein	S
86871c0g1i1	18.4	17.78	0.73	yes*	Uncharacterized protein LOC104224418	S
20225c0g1i1	28.8	17.76	0.18	yes	-	
22899c0g2i1	53.5	17.74	0.28	-	Uncharacterized protein LOC107760627	T
21258c0g1i2	61.9	17.72	0.02	-	Nucleoside diphosphate kinase 3	S
75249c0g1i1	44.0	16.17	-0.23	-	Uncharacterized protein LOC107794475	S
18187c0g2i2	17.8	15.68	0.11	yes	Uncharacterized protein LOC104232736	S
64896c0g1i1	10.3	14.35	0.36	-	-	
26833c0g2i1	24.0	14.24	11.16	-	Ethylene-responsive transcript. factor 4	S
16240c0g1i1	26.9	14.05	0.05	yes	4-coumarate--CoA ligase	S
25958c0g1i2	17.8	12.81	7.69	yes*	Heat stress transcription factor A	S
65623c0g1i1	24.4	12.78	0.59	-	-	
19615c0g1i2	10.6	12.48	0.58	-	Interaptin	T
3660c0g2i1	10.3	12.47	-0.02	-	Diaminopimelate epimerase,	S
30135c0g1i1	15.0	12.36	1.03	-	Cytochrome P450 78A5	S
3660c0g1i1	11.8	12.26	0.10	-	Diaminopimelate epimerase,	S

Table 1 continued

30979c0g1i5	50.1	12.23	1.19	-	GDSL esterase/lipase	S
33333c2g1i1	18.7	11.97	-0.05	-	-	
12543c0g1i1	35.3	11.85	0.28	-	Uncharacterized protein LOC104210975	S
28570c0g1i1	42.9	11.57	0.04	yes	Protein disulfide-isomerase	S
85964c0g1i1	10.5	11.00	0.20	yes*	Heat shock protein 83	S
23580c0g1i1	104.5	11.00	-0.33	-	Early light-induced protein 1, chloroplastic	S
18777c2g1i1	11.4	10.71	0.56	-	Uncharacterized protein LOC107823745	S
26611c0g1i2	29.2	10.55	-0.18	yes	-	
32166c0g3i1	25.4	10.45	0.37	-	Uncharacterized protein LOC107771105	T

Columns 3 and 4 are expressed in terms of expression fold change. Column 5 shows contigs which failed to amplify using PCR in TI 1406; those that also failed to amplify in K326-*va* and TN90 are indicated with an asterisk. Column 7 shows the best matching ancestral tobacco species (S = *N. sylvestris,* T = *N. tomentosiformis*).

There are two notable omissions for the genes listed in the Table 1. First, contig 30706c0g2i3, corresponding to the eIF4E1.S isoform originating from the *N. sylvestris* genome and implicated in PVY susceptibility [11] is absent. Contig 30706c0g2i3 has an expression fold change of 1.95 lower in TI 1406 and 1.33 lower in K326-*va*. As there are numerous eIF4E isoforms in the tobacco genome that share high sequence identity, legitimate expression differences were likely masked for 30706c0g2i3 due to cross-mapping during alignment. Second, contigs annotated as light-harvesting chlorophyll a/b-binding proteins (lhcb), some of which have been reported as missing in *va* [11] (and therefore likely *VAM* as well), encounter the same issue. The lhcb contigs that show the greatest similarity to S genome lhcb genes appear with only minimally higher expression in K326 than K326-*va* or TI 1406. Both S genome lhcbs and 30706c0g2i3 show greater differential expression when processed RNA-seq reads are aligned with no mismatches or gaps allowed. However,

we found imposing this level of stringency on gross alignments was less useful, as false positives arising from the natural polymorphisms that exist between K326 and TI 1406 became overwhelming.

As *N. tabacum* originally arose through an ancient hybridization between individuals most closely related to *N. sylvestris* and *N. tomentosiformis* [20] and each of these species has been sequenced in some depth [21], potential information about genetic ancestry can be gained by looking at which of these two ancestral tobacco species is a better match for a given differentially expressed contig. Provided adequate coverage and annotations exist, higher quality alignments are expected with the *N. sylvestris* genome for genes potentially missing in *va* and *VAM* plants. This is due to: (1) *N. sylvestris* being highly susceptible to potyvirus infection while *N. tomentosiformis* is resistant; (2) extra-cuticular CBT-diols are present on the leaves of *N. sylvestris* and are not found in *N. tomentosiformis* [22]; and (3) the *va* locus has be previously mapped to a portion of chromosome 21 expected to originally have been inherited from *N. sylvestris* [11]. Thirty-five of the contigs shown on Table 1 returned *N. sylvestris* as the top alignment, while nine contigs matched most closely to *N. tomentosiformis* sequences.

Candidate Sequence Validation through PCR and Southern Blot Analysis

We attempted to design primers to amplify each of the contigs found in Table 1 and all known eIF4E isoforms, including the suspected PVY susceptibility isoform eIF4E1.S (30706c0g2i3) (Table A2.1, Applendix). Primers were checked against the complete list of

our K326 contigs and online resources in an effort to improve their specificity. Primers were run with K326, K326-*va*, and TI 1406 genomic DNA isolated from the plants used in RNA-seq analysis. Additionally, DNA from TN90 was included as it also possesses the *va* locus.

Twenty-two of the contigs listed in Table 1 failed to amplify in TI 1406, seven of which also failed to amplify with genomic DNAs isolated from K326-*va* and TN90. Although the other 33 contigs could not be validated by PCR, they are still considered potential candidates since it was their short size, repetitive nature and/or the inability to produce discriminating primers that prevented definitive PCR testing. Representative examples of the PCR analyses are shown in Figure 1. Included among the contigs that were PCR verified are a 3-isopropylmalate dehydratase (35943c0g2i1), a gene that was also identified in the RNA-seq analysis of the *va* locus reported by Julio *et al.* [11], and several contigs encoding CBTS genes (represented by 24507c0g1i1, 22989c1g1i3, 33989c1g2i1, 33989c1g1i2, 33989c1g1i4, 33989c1g1i5, and 33989c1g1i6). The multiple entries annotated as CBTS-2a or CBTS-3 may be due to incomplete or differently spliced *de novo* contig assemblies. In *N. sylvestris*, three CBTS genes, *NsCBTS-2a*, *NsCBTS-2b*, *NsCBTS-3*, and one likely pseudogene, CBTS-4, have been reported [23]. BLAST was used to search for additional contigs with similarity to *N. sylvestris* CBTS sequences within our assembled K326 transcriptome. All potential isoforms revealed in these searches, however, were those already represented in Table 1. To directly test for the presence of CBTS-homologous genes in TI 1406, an amplified portion of the CBTS-2a cDNA was used as a hybridization probe in a Southern blot analysis.

The blot shown in Figure 2 shows that all bands are completely missing in TI 1406, suggesting that all CBTS-like genes are absent in the *VAM* mutant line.

Deletion size estimation through analysis of public genome data

The RNA-seq results support the hypothesis that the *VAM* locus is a deletion mutant encompassing a larger proportion of chromosome 21 than the deletion that defines *va*. To gain insight into the physical sizes of the respective deletions, BLAST was used to align contigs with significant K326/TI 1406 expression changes (>3-fold) to scaffolds of the v4.5 K326 tobacco genome deposited in GenBank by Edwards *et al*. [25]. Alignments were individually inspected, as contigs representing mRNAs may not be entirely contained in a single DNA scaffold or may extend over multiple exons. Nevertheless, a ≥99% similarity was required over aligned sections. Figure 3 displays a representation of the layout of *va/VAM* with an expanded section of the end of the long arm of chromosome 21 shown with scaffolds that have been anchored on the chromosome by Edwards *et al*. [25]. Each of the anchored scaffolds in the region predicted to be encompassed by *VAM* were further validated using PCR (data not shown). Additional steps were taken to fill in gaps between anchored scaffolds. First, Tobacco Infinium-30k markers (https://solgenomics.net/cview/map.pl?map_version_id=178) were aligned to all Edwards *et al*. [25] scaffolds. Positive hits to anchored scaffolds are shown highlighted in light green in Figure 3. Scaffolds containing a marker that mapped in the expected *VAM* deletion region, but have not been anchored to the physical map are shown in Table 2. Finally, a draft genome assembly generated by the

Yunnan Academy of Tobacco Agricultural Sciences for tobacco cultivar RBST was aligned to Edwards *et al*. [25] scaffolds and added to our VAM assembly to bridge gaps in the anchored assembly. These scaffolds, in predicted order, are shown in Table 2 along with their length and the number of genes predicted to be present on the scaffold (v4.5 genome annotations [25]), and whether highly differentially expressed genes from our study were found on the scaffold.

Two *N. tabacum* SSR markers located at the end of chromosome 21 developed by Bindler *et al*. [13], designated PT60946 and PT60057, are reported as being absent in *va* plants [11]. PT60946 and PT60057 along with other markers from [13] in the same region of chromosome 21 were tested in lines K326, K326-*va*, TN90 and TI 1406. In agreement with the findings of Julio *et al*. [11], PT60946 and PT60057 did not amplify in K326-*va* or TN 90. As expected, PT60946 and PT60057 also did not amplify genomic DNA from TI 1406, whereas all other tested markers in this region were PCR positive in all lines.
Using the markers in Table 2 and the presence of genes with significant K326/TI 1406 expression change, all scaffolds between Nitab4.5_0003331 and Nitab4.5_0003487 have the potential to be missing in *VAM*. The markers shown represent a map distance of 32.3cm (Tobacco Infinium-30k). Scaffolds shown above Nitab4.5_0001441 in Table 2 are predicted to be deleted only in *VAM*, and not *va.*

Table 2 Predicted VAM scaffolds sorted by a hybrid physical/genetic mapping order.

Anchored Scaffolds	Scaffolds with markers	RBST ordered Scaffolds	Scaffold length	Marker ID	Map cM	Diff-gene	# of gene
Nitab4.5_ 0003331			322448	-	-	-	15
	0007068		139279	Nt1AC8849 Nt1AE0363	0 2.3	Y*	8
	0008902		97203	Nt1AC4113	1.8	-	3
	0013033		33607	Nt1AC7974	2	Y	2
	0008971		95815	Nt1AD4189	2.3	Y	6
Nitab4.5_ 0001441	0001441	0001441	550039	Nt1AE3136 [a]	4.6	Y*	15
		0003268	320528	-	-	Y*	12
		0010227	72084	-	-	Y	1
		0006402	160683	-	-	Y	1
		0002210	421245	-	-	Y	13
		0002814	355923	-	-	Y	16
		0002507	388355	-	-	Y*	13
		0002458	393533	-	-	Y*	13
		0005928	177860	-	-	-	2
		0004518	237847	-	-	-	1
		0006329	163221	-	-	-	1
		0001037	657976	-	-	-	2
		0008217	112516	-	-	-	1
	0003843	0003843	276436	Nt1AC3736	20.4	-	6
Nitab4.5_ 0002177	0002177	0002177	424490	Nt1AE7112 [a]	21.9	Y*	17
Nitab4.5_ 0003711	0003711	0003711	284557	Nt1AD8309	22.7	Y	7
Nitab4.5_ 0001375	0001375		566801	Nt1AE5740 Nt1AE8790	22.7 23.2	Y*	18
	0003487		300957	Nt1AA4543	32.3	Y*	11

a Markers did not have a perfect match to scaffold due to incomplete scaffold sequencing
Diff-gene: a Y indicates that at least one of our contigs displayed >3-fold K326/TI 1406 expression change aligned to this scaffold; a Y* means contig from Table 1 appears on the scaffold
# of genes: genes predicted by Edwards *et al*. [25] to be within this scaffold (annotations found in Table A2.2, Applendix).

Of particular note, Nitab4.5_0002814 and Nitab4.5_0002210 have multiple chlorophyll a-b binding proteins matching those found to be missing in plants with *va* according to Julio *et al*. [11]. Nitab4.5_0002814 contains eIF4E1.S and Nitab4.5_0002210 contains the PCR verified contigs S27183 and S05079 from [11]. Two other PCR verified genes from their study, S25284 and S14740, likely reside on Nitab4.5_0001375. The CBTS gene family seems to be poorly assembled in current genomic resources, with the 1453bp-long Nitab4.5_0221271 being the closest match for CBTS-2a, and no other high quality scaffold hits for other CBTS genes are found in publically available databases.

The total scaffold length for the v4.5 *N. tabacum* genome scaffolds shown in Table 2 is 6.55Mbp, where ~1Mbp are exclusively missing in *VAM* and ~5.5Mbp are absent in both *VAM* and *va*. Based on the v4.5 genome annotations, the scaffolds in Table 2 are predicted to contain 184 genes. Individual gene annotations are displayed in Table A2.2, Appendix. The K326 genome scaffolds deposited in GenBank by Sierro *et al*. [2] (assembly: GCA_000715075.1) were also investigated for their potential to expand our knowledge of *VAM,* but were ultimately not included because they did not substantially add to or alter the model of *VAM* and *va* structure shown in Fig. 3 and Table 2.

Localizing a *va* end point.

If one could identify an exact deletion junction for the *va* locus, it should be possible to create co-dominant markers to assist in the breeding of this important recessive locus into additional tobacco varieties. As an attempt to define such a junction, primers were

designed along the length of the anchored genomic scaffolds expected to be absent or at least partially missing in *va* (primers listed in Table A2.1, Appendix). PCR results predicted that the terminal end of *va* lies within scaffold Nitab4.5_0001441 (Figure 4). Alignments to the pre-existing TN90 tobacco genomic assembly (Sierro *et al.* [2]) failed to show the exact *va* end point due to a lack of coverage in the particular area of interest. By generating and testing a series of additional PCR-based markers generated across against Nitab4.5_0001441, we were able to localize the likely *va* end point to a 637bp region (nucleotides 306,892 - 307,529 of Nitab4.5_0001441). Although 637bp is a short enough distance for determining the exact deletion junction through genome walking, multiple attempts to walk through the area were only successful in the K326 control; no legitimate genome walking products were obtained from genomic DNAs isolated from either TN90 or K326-*va* (data not shown).

Transformation of K326-*va* and TI 1406 with eIF4E1.S

Two previous reports have shown that the absence of the single, specific eIF4E1.S isoform of the large eIF4E gene family is highly correlated with PVY susceptibility in tobacco [2, 11]. The most compelling evidence that eIF4E1.S non-function is causally responsible for the PVY resistance phenotype in plants possessing *va* (as opposed to other genes located within the deleted chromosomal region) comes from the observation by Julio *et al*. [11] that tobacco plants carrying an ethylmethane sulfonate-induced premature stop codon in eIF4E1.S showed a PVY resistance phenotype similar to plants with *va*. If the lack of a

functional eIF4E1.S gene is indeed responsible for resistance to potyvirus infection, then it would be predicted that the genetic complementation of tobacco lines possessing *va* or *VAM* with a normal copy of the gene would restore a susceptibility phenotype. To test this, a wild type eIF4E1.S cDNA was cloned into a plant expression vector under the transcriptional control of the CaMV 35S promoter and introduced into TN86, which was the original tobacco variety developed containing the *va* gene [8]. As show in Figure 5A, quantitative RT-PCR analysis revealed several T_0 transgenic events in which eIF4E1.S was successfully expressed. Inoculation of the eIF4E1.S-transformed TN86 plants with PVY^{ZT-5}, an isolate of PVY found in Southern China, resulted in a chlorotic mosaic leaf phenotype and stunted growth. In contrast, TN86 plants transformed with the control vector showed no symptoms of virus infection. Examples of each genotype 21 days post-inoculation are shown in Figures 5B and C. Biochemical evidence of a robust PVY infection in the non-inoculated leaves of the eIF4E1.S-expressing plants was obtained by ELISA and Immunostrip assays using antibodies directed against the PVY coat protein (Table 2.3, Appendix).

In addition to TN86, 35S::eIF4E1.S constructs were introduced into TI 1406 and K326-*va*. Some T_0 individuals were inoculated with TEV, and others were inoculated with a particularly necrotic strain of PVY (PVY^{NN}). Similar to the results obtained in the TN86 background, symptoms of virus infection were readily observed in TI 1406 and K326-*va* plants transformed with the eIF4E1.S transgene and not in vector only controls (Figure A2.1, Appendix). These results support the conclusion that the re-introduction of eiF4E1.S gene

into tobacco plants containing the *va* or *VAM* deletion mutations is sufficient for restoring potyvirus susceptibility to these plants.

Investigation of genes involved in diterpene synthesis in TI 1406

As described in Background, previous studies of TI 1406 have concluded that the failure of plants possessing the *VAM* mutation to produce leaf surface diterpenes (as well as the sucrose ester BMVSE) is due to a lack of functional chloroplasts in the heads of glandular trichomes in these plants [7, 14, 29], as opposed to the alternative model that the genomes of these plants simply fail to have the genes/enzymes required to synthesize these compounds. This is best exemplified by the study by Nielson and Severson [14] which showed that the genes needed for Z-abienol and BMVSE production were indeed present in the TI 1406 genome, but could not be functionally manifest unless separated from the trichome deficiency-inducing *VAM* locus (called *te* in that paper). Figure 6A shows typical examples of K326 and TI 1406 trichome heads under a light microscope (60X magnification). One of the most striking results of our RNA-seq (and Southern blot) analysis of TI 1406 was the absence of CBTS genes in this line (Table 1; Fig. 2). Within the context of the current model of *VAM* locus function, there are two possible explanations of the relationship between the lack of CBTS genes in TI 1406 and its defective trichome secretory phenotype: (1) the production of α-and β–CBT-diols via CBTS activity is required for the formation of functional chloroplasts in the heads of secretory trichomes in tobacco plants; or (2) the deletion of CBTS genes in *VAM* plants is “coincidental” to the defective trichome

phenotype, with another gene(s) in the deletion region being responsible for the trichome-specific chloroplast developmental defect. In the first scenario the restoration of CBTS gene activity would be predicted to complement the trichome secretory defects of TI 1406 plants, whereas in the second scenario it would not, and α-and β–CBT-diol production would still not be observed even after reintroduction of a functional CBTS gene.

As an initial step in exploring this phenomenon, a comparison was made between expression levels of the genes unique to CBT-diol and Z-abienol production by aligning the RNA-seq reads from K326 and TI 1406 to the assembled K326 transcriptome. *CYP71D16* encodes a cytochrome P450 that is expressed specifically in glandular trichomes and functions after the CBTS step by catalyzing the conversion of CBT-ols to CBT-diols [18]. *NtCPS2*, encoding a class-II terpene synthase, and *NtABS*, encoding a kaurene synthase-like protein, are the structural genes responsible for the production of Z-abienol from GGPP [27]. Expression differences between K326, K326-*va* and TI 1406 for this group of diterpene-associated genes are shown in Table 3. Among these, only contigs corresponding to CBTS appear to be differentially expressed. Furthermore, visual inspection of the alignments of the reads corresponding to *NtCPS2, NtABS,* and *CYP71D16* from TI 1406 revealed no polymorphisms that would suggest that the encoded enzymes would not be functional (data not shown). These results are consistent with the hypothesis proposed by Nielson and Severson that TI 1406 individuals possess a functional Z-abienol pathway, despite its lack of production in these plants [14]. We also conclude that the *CYP71D16* gene that is responsible for conversion of α- and β-CBT-ols to α-and β–CBT-diols is also fully functional.

Table 3 Select tobacco diterpenoid pathway genes.

Gene	Best contig match*	K326/K326-va fold change	K326/TI 1406 fold change
CYP71D16	32566c0g1i2	-0.09	0.68
NtCPS2	20163c0g2i1	0.34	0.14
NtABS	34936c0g2i11	0.02	-0.15
NtCBTS-2a	33989c1g1i6	0.06	1443.83
NtCBTS-3	33989c1g1i3	0.18	2351.5

*The contig from our K326 transcriptome with the best match with sequences entered in GenBank.

To test whether the expression of CBTS activity in TI 1406 is capable of restoring the synthesis of □-and □-CBT-diols in this background, a full length version of *NtCBTS-2a* was PCR amplified from K326 using primers (cembBam2_F and cembSac2_R, Table 2.1, Appendix) designed based on the *NsCBTS-2a* sequence of *N. sylvestris* [23]. BLASTN alignments confirmed that *NtCBTS-2a* is the same CBTS gene as the one designated *CYC-1* by Wang and Wagner [16] who generated RNAi constructs against its sequence to demonstrate that down regulation of this gene effectively inhibited α- and β-CBT-diol production in tobacco. Two promoters were used to drive expression of *NtCBTS-2a*, the 35S promoter of CaMV and the native *NtCBTS-2a* promoter (designated CEMBpro). The latter promoter sequence was obtained by amplifying a 1,812bp region upstream of the *NtCBTS-2a* initiation of translation site. Positions -1 through -988 (in relation to the start ATG) of this 1,812bp sequence are >99% identical to the 988 bp region of the *Nicotiana sylvestris NsCBTS-2a* gene promoter that was previously shown to confer a high level of trichome-specific gene expression (Genbank ID HM241147.1)[23].

A minimum of 10 independent T_0 individuals were generated using 35S::CBTS-2a, CEMBpro::CBTS-2a and vector-only constructs in both TI 1406 and K326 backgrounds. Chemical analysis of leaf surface washes of all T_0 plants generated in this study failed to reveal a single TI 1406 plant that accumulated any more than negligible amounts of α- and β-CBT-diols, regardless of the promoter used. These results, combined across all T_0 individuals in each genotypic class are shown in Figure 6B. In contrast, all transgenic plants in this study in the K326 background showed substantial α- and β-CBT-diol accumulation on their leaf surfaces. Interestingly, the introduction of 35S::CBTS-2a and CEMBpro::CBTS-2a constructs in K326 did not lead to α- and β-CBT-diol accumulation levels that were significantly different from that observed in the vector-only controls (Figure 6B), suggesting that the endogenous CBTS activity in K326 is already quantitatively metabolizing the available GGPP substrate. Given that the reintroduction of *NtCBTS-2a* in TI 1406 did not alter the leaf surface chemistry phenotype of plant, it was not surprising that the chloroplast-deficient phenotype of the trichome heads also remained unchanged (Figure 6C).

DISCUSSION

Nature and composition of *va* and *VAM*

In this study, we investigated the chromosomal regions associated with the *va* and *VAM* deletion mutations. Previous studies have shed light on certain specific genes

exclusive to *va* and their location within the tobacco genome [2, 11]. However, the magnitude of the deletion that defines *va,* as well as its relationship to the larger *VAM* deletion from which *va* was believed to have been derived were unknown. Alignment of RNA-seq reads to a *de novo* assembled K326 transcriptome led to the identification of 55 contigs whose expression levels were at least 10-fold lower in TI 1406 (Table 1).

In creating this differentially expressed contig list, steps were taken to bring it to a high standard. The high potential for false positives, due to sequence or sequencing differences in homologous contigs of K326 versus TI 1406 was reduced by utilizing publically available K326 RNA-seq data to identify and remove contigs with altered expression that may have been independent of the *va* or *VAM* loci. Additionally, requiring that the contigs have close sequence similarity between our K326 transcriptome and an assembled TN 90 transcriptome reduced the impact of inherent TI 1406/K326 differences. As K326-*va* and K326 are near-isogenic lines and the origin of TI 1406 is not clear, there will inevitably be more false positives in the K326/TI 1406 expression comparisons. Twenty-two contigs from Table 1 were validated by PCR as not being able to be detected in tobacco genomes containing the *VAM* or *va* mutations, 15 of which were exclusively missing in *VAM*. Due to our inability to develop suitable PCR primers that could specifically discriminate the remaining 33 contigs of Table 1, we are unable to conclude which of these may also be represented within the chromosomal regions deleted in *VAM/va* as opposed to being localized elsewhere in the genome and being differentially expressed as a secondary consequence of the deletion on chromosome 21.

Through aligning markers and genes from Table 1 to publicly available tobacco genome scaffolds and later amplifying targeted regions of these scaffolds, we were able to define a set of anchored and unanchored scaffolds which are likely to be missing in *va* and/or *VAM* plants. Additional genome assembly data developed at the Yunnan Academy of Tobacco Sciences was used to define the order of numerous unanchored scaffolds to ones that have been anchored (Table 2). Insights into scaffold order were also gained by identifying unanchored scaffolds that contained previously published markers, either SSR [13] or Tobacco Infinium-30k [25], which had been mapped to this region of chromosome 21. From these collective analyses, the *VAM* deletion is estimated to encompass a minimum of 24 scaffolds representing 6.55 Mbp and 184 genes. Based on the markers and map distances calculated, *VAM* would represent a deleted area encompassing at least 32.3 cM (of a 123.4 cM total of chromosome 21 estimated using Tobacco Infinium-30k markers). The predicted *va* deletion spans the majority of that represented by *VAM*, including 18 scaffolds over a 5.5 Mbp region containing 143 genes with a genetic distance of ~27.7 cM.

Scaffold Nitab4.5_0003651 is the last anchored scaffold on the *VAM*-containing long arm of chromosome 21 (Fig. 3). Genes on this scaffold showed slightly higher expression in K326 than TI 1406. Attempts to PCR validate sections of the scaffold resulted in amplification in all backgrounds. However, as the genes in this area may be repetitive, masking expression differences and the hindering the ability to design differential PCR primers, we cannot eliminate the possibility that the *VAM* deletion extends all the way to the end of the chromosome.

As a caveat to the proposed structure and content of the *VAM* and *va* deletion mutations presented here, we acknowledge that they were generated under the assumption that the events leading to creation of *VAM* and *va* resulted in "clean" deletions, whereby a large contiguous chromosomal region of chromosome 21 was removed. It is entirely possible that these deletion mutations are more complex, representing the elimination of certain portions of the native chromosome within this region while retaining others, possibly in rearranged form or order. Thus, the structures and data presented in Figure 3 and Table 2 should be viewed as a working model, with the understanding that the true nature of *va* and *VAM* will only be fully understood by whole genome sequence analysis of lines possessing these loci.

Whole genome sequence information would also be useful in identifying specific breakpoints at the two deletion junctions (and possibly more if the nature of the deletion is complex) in wild type versus *va* individuals. The *va* locus has been introduced into several commercial tobacco varieties as a valuable source of potyvirus resistance. The transfer of *va* to new cultivars, however, is not simple due to its recessive phenotype and the fact that all molecular markers directed toward this locus to date are dominant, and thus cannot distinguish homozygous dominant individuals (*VA/VA*) from heterozygous carriers (*VA/va*) in segregating populations. Knowledge of the sequence spanning a deletion break in genomes with *va* could enable the development of PCR-based markers capable of uniquely amplify *va*, which when coupled with primers corresponding to the wild type sequence at the breakpoint junction could be used as a co-dominant markers and greatly facilitate the

breeding of this important trait into new varieties. Despite our numerous attempts at genome walking, however, we were unable to identify the location of the deletion breakpoint within a 637bp region of scaffold Nitab4.5_0001441.

Exploration of CBTS and trichome phenotypes in *VAM*

Trichomes of TI 1406 individuals lack chloroplasts, and instead possess unusual structures described as membrane-bound or free inclusions [7]. Our results suggest that genes involved in the *cis*-abienol pathway and the *CYP71D16* gene of the CBT-diol pathway are expressed at normal amounts in TI 1406. Furthermore, the sequences of these genes in TI 1406 possess no mutations, which strongly suggests that these steps of in diterpene synthesis in TI 1406 are fully functional. In contrast, contigs representing members of the CBST family appeared as the most differentially expressed in our RNA-seq analysis, and Southern blot assays failed to show any hybridization to CBTS-homologous sequences in TI 1406, using *NtCBTS-2a* as a hybridization probe. Our collective data presents compelling evidence that a small CBTS gene family (comprised as a minimum of *NtCBTS-2a* and *NtCBTS-3*) is found on the region of chromosome 21 that is deleted in plants possessing the *VAM* locus, but not *va.* The failure to directly assign any CBTS gene to this region, however, is likely attributable to the fact that neither of the two large publically available tobacco genome sequencing initiatives could place CBTS-homologous sequences on any scaffold greater than 2640 bp in size (AWOJ01S331035.1). Thus, it is likely that sequences in the vicinity of the CBTS gene family are recalcitrant to sequencing and/or assembly by the

approaches used by both groups. Support for the CBTS gene family residing on chromosome 21 also comes from Southern blotting assays conducted using genomic DNA from the line Red Russian Null E (a line nullisomic for chromosome E, which is the older nomenclature for chromosome 21). Like the TI 1406 Southern blots, no hybridization to *NtCBTS-2a* probes could be detected in Red Russian Null E, providing additional evidence that the CBTS gene family resides on this chromosome (data not shown).

The expression of a *NtCBTS-2a* transgene in TI 1406 under the transcriptional control of either the 35S promoter, or its own native promoter, failed to restore CBT-diol production in these plants. *Z*-abienol was also not detected in any plant (data not shown), despite there being a functional pathway for its synthesis in TI 1406. Kandra *et al*. [29] showed that diterpene biosynthesis in tobacco accession TI 1068 is dependent on light, and can be inhibited by the photosynthesis inhibitor DCMU. Because of these observations, the authors proposed that the biosynthesis of leaf surface exudates from the heads of glandular trichomes in tobacco is dependent on energy and/or metabolites provided by the chloroplasts in these cells, a notion that was later supported by comparative ultrastructural analysis of the trichome heads of TI 1068 versus TI 1406 [7]. Our results suggest that a gene(s) other than *NtCBST-2a* that lies within the region of the *VAM* deletion that is not shared with *va* is responsible for the development of functional chloroplasts within the heads of secretory trichomes in tobacco. Our current model identifies 41 genes that reside within the ~1 Mbp region exclusively missing in *VAM* mutants, and thus represents the list of candidates for enabling normal chloroplast development within the trichome (though

undoubtedly more genes will be added as the whole tobacco genome sequence becomes further refined and more complete). Included in this region are several potential transcription factors (Table A2.2, Appendix). Furthermore, although multiple lhcb genes have been shown to be missing in *va* [2], an additional lhcb is located on a scaffold expected to be part of *VAM* and not *va* (Nitab4.5_0007068g0040.1 from Table A2.2, Appendix; 23580c0g1i1 from Table 1). The effect that the absence of lhcb genes may have on the *va* or *VAM* phenotypes is unknown, but the loss of a requisite trichome-specific lhcb gene may represent one mechanism whereby chloroplast dysfunction within the trichome head could become manifest.

CONCLUSIONS

The results presented here provide an in depth characterization of the deletion mutants *va* and *VAM*, two loci of both historical and commercial importance in tobacco. By coupling results obtained from RNA-seq analyses with publically available whole genome sequence information, we were able to develop an initial physical map of the region of chromosome 21 that is deleted in tobacco plants containing the *va* or *VAM* loci. The characterization of select genes within these deletion mutants lead to confirmation of the role of *eiF4E1.S* in facilitating susceptibility to potyvirus infection, as well as the discovery that the small CBTS gene family involved in the biosynthesis of the major class of diterpenes found on the leaf surface of tobacco appears to be missing in plants possessing *VAM.*

Finally, although the majority of the chromosomal deletion that defines *VAM* is also deleted in *va*, within the deletion region unique to *VAM* are one or more genes responsible for the production of functional chloroplasts within the trichome head. The results of this study provide the foundation for further investigation of the mechanisms by which viable chloroplasts are maintained within the heads of glandular trichomes.

MATERIALS AND METHODS

Plant growth conditions

Tobacco plants for all experiments were grown at room temperature using a 16h light, 8h dark cycle. Plants transformed with various *NtCBTS-2a* or vector control constructs were transferred to a greenhouse at approximately the eight-leaf stage and grown to maturity prior to analysis of leaf surface diterpene content.

Sequencing and transcriptome assembly

Total RNA was isolated from 100mg of pooled young leaf tissue of tobacco lines K326, K326-*va* and TI 1406 using the Trizol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol. RNA quality was established using a 2100 Bioanalyzer (Agilent Technologies). HiSeq sample preparation was completed using the TruSeq RNA Library Prep Kit v2 (Illumina Inc. San Diego, CA, USA). All three samples were multiplexed on a single lane of an Illumina HiSeq chip for 100bp paired-end sequencing. For certain applications, the

K326 RNA-seq reads deposited in GenBank by Sierro *et al.* (SRR955772) [2] were also downloaded and utilized.

Processing of raw sequencing reads was performed on all four RNA-seq sets with the Trimmomatic tool (v0.32)[30] using the options: LEADING:13 TRAILING:13 SLIDINGWINDOW:3:21 MINLEN:55. The Trinity *de novo* assembly package (v2.0.6) was used to assemble a transcriptome using only our processed K326 reads [31, 32]. Contigs under 500bp in length were removed as well as contigs with under a 10-fold coverage.

Aligning reads back to transcriptome for deletion discovery

To determine differentially expressed contigs, the processed reads from K326, K326-SRR955772, K326-*va* or TI 1406 were mapped back to the assembled K326 transcriptome using BWA-MEM (v0.7.12) with the options: -c 60000 [33]. Samtools Idxstats function was used to collect data on alignments.

PCR and Southern blot analysis

PCR amplification of contigs from Table 1 and amplification done to verify and examine scaffolds was performed with primers listed in Table A2.1, Appendix. *Taq* DNA Polymerase (New England BioLabs, Ipswich, MA) was used with thermal cycling conditions consisting of an initial denaturing of 95°C for 30s; 32 cycles of denaturation at 95°C for 10s,

annealing at 60°C for 20s, and extension at 68°C for 1m and a final cycle of extension at 68°C for 5m. Quantitative real-time PCR was conducted as previously described [36].

Vector construction and plant transformation

Cloning the entire cDNA coding region of *eiF4E1.S* from K326 (corresponding to GenBank accession KF155696) was followed by verification via DNA sequencing. Primers eIF4E_C_F and eIF4E_C_R containing BamHI and SacI sites, respectively, at their 5' ends were used to amplify the complete open reading frame. PCR products were digested with BamHI and SacI and ligated into plant transformation vector pBI121 that had been digested with the same enzymes (replacing the GUS gene with *eiF4E1.S*) to produce construct 35S:: eiF4E1.S. A full-length *NtCBTS-2a* cDNA and its corresponding promoter were amplified using primers CBTS2aF/CBTS2aR and cembPro_F/cembPro_R, respectively (again placing BamHI and SacI sites on their respective ends). The 35S::CBTS-2a construct was generated by replacing the GUS gene in pBI121 with the *NtCBTS-2a* cDNA using BamHI and SacI digestion and ligation as described for 35S::eiF4E1.S. The *NtCBTS-2a* promoter (CEMBpro) was isolated by amplifying an 1,812bp fragment immediately upstream of the ATG start codon. The CEMBpro::GUS construct was generated by replacing the GUS gene in pCAMBIA-1391 using SalI and BamHI sites. CEMBpro::CBTS-2a was made using CEMBpro and CBTS-2a linked at a BamHI site and integrated into pCAMBIA-1390 at SalI and EcoRI sites. Primers mentioned here are provided in Table A2.1, Appendix. All vectors described in

this study were subsequently transformed into *Agrobacterium tumefaciens* GV3101 for plant later transformation using the freeze/thaw shock method [35].

PVY and TEV viral inoculations

Potyvirus strains were maintained on susceptible tobacco cultivars in insect-proof cages. Strain PVY^{ZT-5} was isolated from a tobacco field in Yunnan Province, China and is maintained by the Yunnan Academy of Tobacco Agricultural Sciences. TEV and nectrotic PVY strain PVY^{NN} are maintained at North Carolina State University. Virus inoculum was prepared by macerating systemically infected leaf tissue in phosphate buffer using a mortar and pestle. Approximately 1% (w/v) of quartz sand (200 mesh) was added to the inoculum and filtered through a 40 mesh Nylon net. Plants were inoculated at approximately the 7-8 leaf stage. Inoculum was applied to two leaves per plant using a high-pressure spray gun. Plants were evaluated either 14 or 21 days after inoculation.

GC-MS analysis

Trichome exudate analysis was performed using GC-MS with minor changes as described by Vontimitta *et al*. [35]. Briefly, ten 1.5cm leaf punches per plant were collected and pooled. Samples were immediately placed on ice and remained there until analysis. Leaf surface exudates were collected by washing the leaves twice with CH_2Cl_2, adding Na_2SO_4, and incubating overnight. An Agilent HP 6890 GC-FID (Santa Clara, CA) was used for gas chromatographic analysis with a helium carrier.

REFERENCES

1. Gur A, Zamir D. Unused natural variation can lift yield barriers in plant breeding. PLOS Biol. 2004;2:1610–1615.
2. Sierro N, *et al*. The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun. 2014;5:3833.
3. Koelle G. Genetic analyse einer Y-virus (Rippenbraune) resistenten mutante der tabaksorte Virgin A. Zuchter. 1961;31:71–71.
4. Johnson MC, Pirone TP. Evaluation of tobacco introduction 1406 as a source of virus resistance. Phytopathology. 1982;72:68-71.
5. Fischer DB, Rufty RC. Inheritance of partial resistance to tobacco etch virus and tobacco vein mottling virus in burley tobacco cultivar Sota 6505. Plant Dis. 1993;77:662-666.
6. Nielsen MT, Jones GA, Collins GB. Inheritance pattern for secreting and nonsecreting glandular trichomes in tobacco. Crop Sci. 1982;22:1051–1053.
7. Nielsen MT, Akers CP, Jarlford VE, Wagner GJ, Berger S. Comparative ultrastructural features of secreting and non-secreting glandular trichomes of two genotypes of *N. tabacum*. Bot Gaz. 1991;152:13-22.
8. Miller RD. Registration of TN 86 burley tobacco. Crop Sci. 1987;25:2,365-366.
9. Noguchi S, Tajima T, Yamamoto Y, Ohno T, Kubo T. Deletion of a large genomic segment in tobacco varieties that are resistant to potato virus Y (PVY). Mol Gen Genet. 1991;262:822–829.

10. Acosta-Leal R, Xiong Z. Complementary functions of two recessive *R*-genes determine resistance durability of tobacco 'Virgin A Mutant' (VAM) to *Potato virus Y*. Virology. 2008;379(2):275–283.
11. Julio E, Cotucheau J, Decorps C, Volpatti R, Sentenac C, Candresse T, de Borne FD. A eukaryotic translation initiation factor 4E (eIF4E) is responsible for the "va" tobacco recessive resistance to potyviruses. Plant Mol Biol Rep. 2015;33:609–623.
12. Truniger V, Aranda MA. Recessive resistance to plant viruses. Adv Virus Res. 2009;75:119–159.
13. Bindler G, Plieske J, Bakaher N, Gunduz I, Ivanov N, Van der Hoeven R, Ganal M, Donini P. A high density genetic map of tobacco (*Nicotiana tabacum L*.) obtained from large scale microsatellite marker development. Theor Appl Genet. 2011;123:219–230.
14. Nielsen MT, Severson RF. Variation for flavor components on leaf surfaces of tobacco genotypes differing in trichome density. J Agric. Food Chem. 1990;38:467-471.
15. Severson RF, Johnson AW, Jackson DM. Cuticular constituents of tobacco: factors affecting their production and their role in insect and disease resistance and smoke quality. Rec Adv Tobacco Sci. 1985;11:105-174.
16. Wang E, Wagner GJ. Elucidation of the functions of genes central to diterpene metabolism in tobacco trichomes using posttranscriptional gene silencing. Planta. 2003;216:686–691.

17. Wang EM, Gan SS, Wagner GJ. Isolation and characterization of the CYP71D16 trichome-specific promoter from *Nicotiana tabacum* L. J Exp Bot. 2002;53:1891–1897.

18. Wang E, Wang R, Deparasis J, Loughrin JH, Gan S, Wagner GJ. Suppression of a P450 hydroxylase gene in plant trichome glands enhances natural-product-based aphid resistance. Nature Biotechnology. 2001;19:371–374.

19. Fricano A, Bakaher N, Corvo MD, Piffanelli P, Donini P, *et al*. Molecular diversity, population structure, and linkage disequilibrium in a worldwide collection of tobacco (*Nicotiana tabacum L*.) germplasm. BMC Genet. 2012;13:18.10.1186/1471-2156-13-18.

20. Leitch IJ, *et al*. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). Ann Bot (Lond). 2008;101,805–814.

21. Sierro N, *et al*. Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. Genome Biol. 2013;14:R60.

22. Severson RF, Jackson DM, Johnson AW, Sisson VA, Stephenson MG. Ovipositional behavior of tobacco budworm and tobacco hornworm: effects of cuticular components from *Nicotiana* species. ACS Symp. 1991;449:264-277.

23. Ennajdaoui H, Vachon G, Giacalone C, Besse I, Sallaud C, Herzog M, Tissier A. Trichome specific expression of the tobacco (*Nicotiana sylvestris*) cembratrien-ol synthase genes is controlled by both activating and repressing cis-regions. Plant Mol Biol. 2010;73:673–685.

25. Edwards KD, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans AD, Bombarely A, Allen F, Hurst R, White B, Kernodle SP, Bromley JR, Sanchez-Tamburrino JP, Lewis RS, Mueller LA.. A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. BMC Genomics. 2017;18:448

27. Sallaud C, Giacalone C, Töpfer R, Goepfert S, Bakaher N, Rösti S, Tissier A. Characterization of two genes for the biosynthesis of the labdane diterpene *Z*-abienol in tobacco (*Nicotiana tabacum*) glandular trichomes. Plant J. 2012;72:1–17.

29. Kandra L, Wagner GJ. Studies of the site and mode of biosynthesis of tobacco trichome exudate components. Arch Biochem Biophys. 1988;265:425-432.

30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–2120.

31. Grabherr MG, *et al*. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnol. 2011;29:644–652.

32. Haas BJ, *et al*. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494–1512.

33. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;arXiv preprint arXiv:1303.3997.

34. Sato, M., Komari, T. and Asaine, K. 1982. Varietal differences in the composition of leaf surface diterpenoids in tobacco. Iwata Tob. Exp. Stn. Bull. 14: 59-71.

35. Sparkes IA, Runions J, Kearns A, Hawes C. Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. Nature Protocols. 2006;1:2019-2025.

36. Shi J, Li W, Gao Y, Wang B, Li Y Song, Z. 2017. Enhanced rutin accumulation in tobacco leaves by overexpressing the *NtFLS2* gene. Biosci. Biotech. Biochem. 81: 1721-1725.

FIGURES

Figure. 1

PCR validation of genes representing the two classes of contigs revealed by RNA-seq analysis. PCR analysis of contigs 35943c0g2i1, 22534c0g3i1 and 86871c0g1i1 shows the typical pattern of genes that are absent in lines containing either *va* (K326 va and TN 90) or *VAM* (TI 1406). Contigs that only fail to amplify in plants with the *VAM* mutation are exemplified by 33989c1g1i6 (CBTS-2a) and 23580c0g1i1.

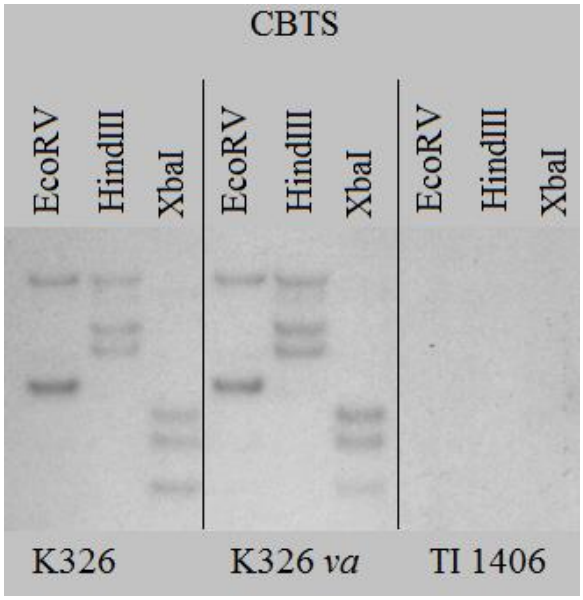


Figure. 2

Southern blot analysis using the *NtCBTS-2a* cDNA as hybridization probe. Genomic DNAs from tobacco lines K326, K326-*va* and TI 1406 were digested with EcoRV, HindIII, and XbaI and analyzed on the same gel.

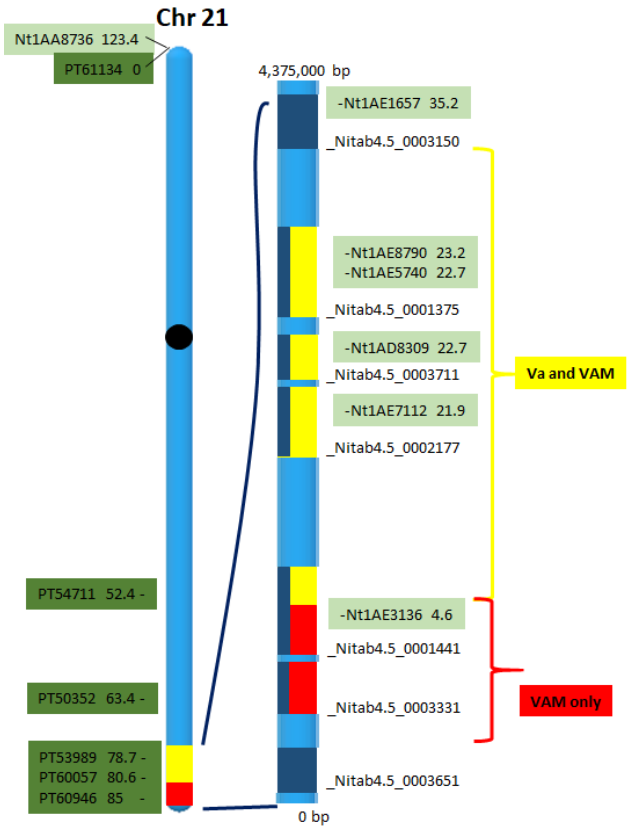


Figure. 3

Map of tobacco chromosome 21, highlighting the region encompassing the *VAM* and *va* loci. Anchored scaffolds missing in both *VAM* and *va* plants are shown in yellow and those uniquely missing in *VAM* in red. Infinium markers used by Edwards *et al*. [25] and aligned to scaffolds are in light green along with their relative distance in cM. Markers described by Bindler *et al*. [13] are dark green along with their predicted distance in cM. Note that the two marker systems begin at different ends of the chromosome and differ substantially in overall predicted map distance.

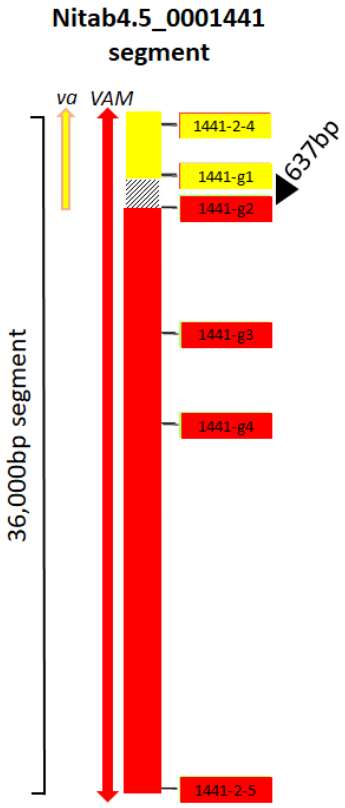


Figure. 4

A 36,000 bp region of scaffold Nitab4.5_0001441 including a proposed *va* deletion junction. Yellow and red highlighted numbers represent PCR markers used to localize one end of the *va* deletion region. Markers highlighted in red score positive in both wild type plants and plants with *va*, and are negative in TI 1406; markers shown in yellow fail to amplify a product in plants containing either *va* or *VAM*. The 637 bp region separating markers 1441-g1 and 1441-g2 (not drawn to scale) is highlighted.

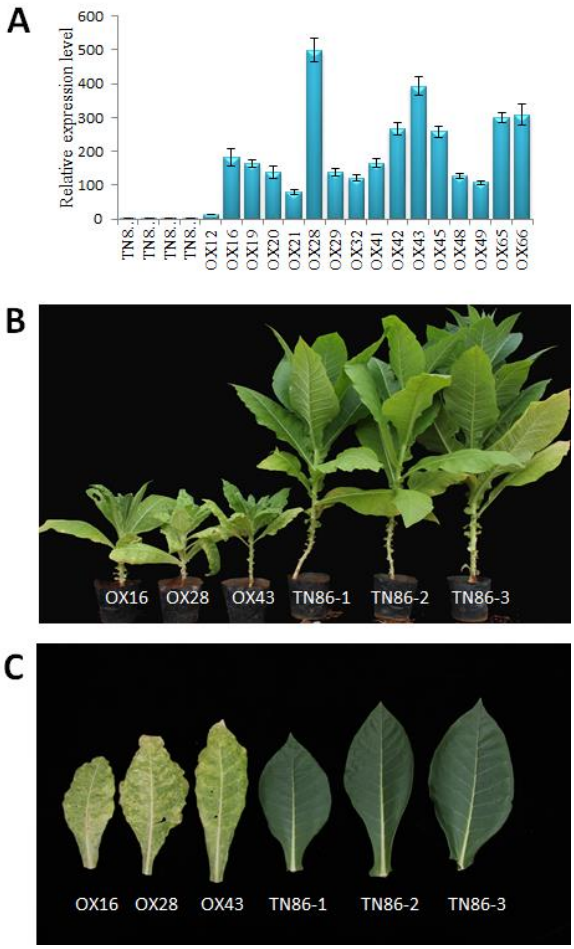


Figure. 5

Expression of eIF4E1.S in TN86 restores susceptibility to PVY. **A** Relative expression of *eIF4E1.S* (in comparison to an endogenous actin gene) of four vector control and 16 35S::eIF4E1.S T_0 transgenic plants using qPCR. Whole plant (**B**) and individual leaf (**C**) phenotypes from three independent T_0 35S::eIF4E1.S (OX) and vector control (TN86) plants are shown 21 days post-infection with strain PVY^{ZT-5}.

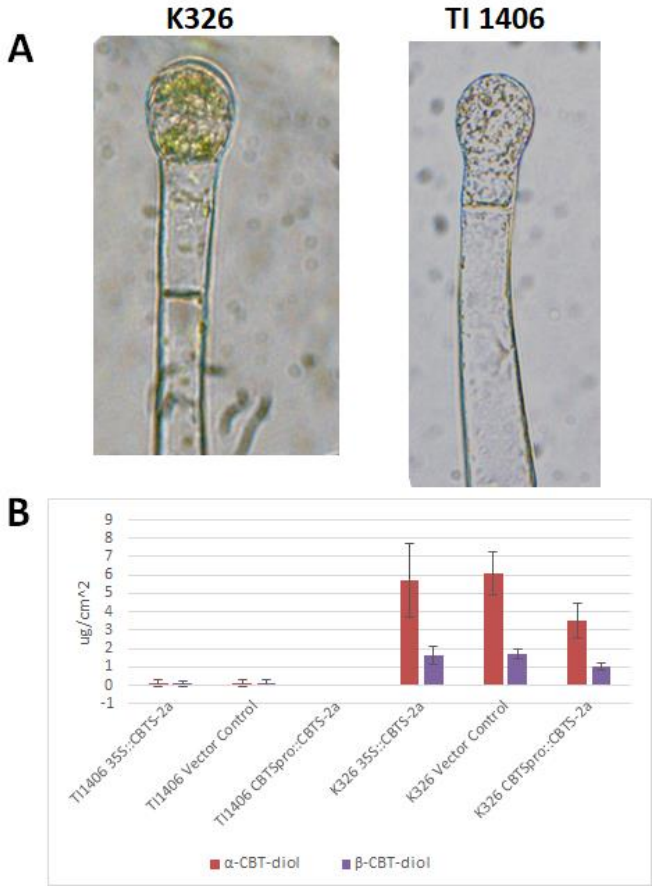


Figure. 6

Ectopic expression of *NtCBTS-2a* fails to restore the ability of TI 1406 trichomes to produce α-and β–CBT-diols. **A** Light microscopy images of typical K326 and TI 1406 trichomes. **B** α- and β–CBT-diol content of TI 1406 plants transformed with 35S::CBTS-2a (n=15), Vector Control (n=10) and CEMBpro::CBST-2a (n=15); and K326 plants transformed with 35::CBTS-2a (n=10), Vector Control (n=10) or CEMBpro::CBST-2a (n=10). Confidence intervals (95%) are provided.

CHAPTER 3

ABSTRACT

Background

Due to advances in genomics technologies, characterizing genomic introgression fragments derived from wild species and introduced into cultivated crops has become increasingly possible. In tobacco, one such introgression fragment, originating from *N. debneyi,* provides a level of resistance to the oomycete *Peronospora tabacina* (blue mold). The blue mold resistance (*bmr*)-containing introgression fragment, however, displays linkage drag with respect to yield when in the homozygous state. In this study, we used RNA-seq technologies and recent tobacco reference genome sequence data to fine map and characterize the introgression fragment containing blue mold resistance.

Results

NC 775 bmr/bmr is a near-isogenic line of cultivar NC 775 that contains the *N. debneyi bmr* introgression fragment via backcrossing using the commercial tobacco variety NC 2000 as the source for *bmr*. RNA-seq was used to analyze genotypes NC 775, NC 775 bmr/bmr, and *N. debneyi* to produce a list of contigs differentially expressed and/or present in NC 775 bmr/bmr and *N. debneyi*, versus the NC 775 control. 25 contigs from this list were verified by PCR as originating from *N. debneyi*. By genotyping the progeny of a NC 775 X NC 775 bmr/bmr cross using sequence characterized amplified region (SCAR) markers reported as flanking the *bmr* locus, 14 recombinant lines were recovered, each possessing differing amounts of *N. debneyi*-derived chromosomal DNA. PCR-based markers were generated

specific to the 26 *N. debneyi* contigs enabling their mapping using the recombinant lines. The results of field trials conducted during the course of this project unexpectedly showed less blue mold resistance in NC 775 bmr/bmr compared to NC 2000. RNA-seq was performed on NC 2000 to determine whether the enhanced resistance in this line could be attributable to additional sequences of *N. debneyi* origin that may have been lost during the backcrossing of the introgression fragment into NC 775, versus the possibility that NC 2000 contains additional blue mold resistant QTLs that were lost in the process. The failure to identify additional *N. debneyi* contigs unique to NC 2000 favored that latter model. Although attempts to field test the 14 recombinant lines for blue mold resistance were unsuccessful, it was discovered that the *N. debneyi*-derived introgression fragment had replaced a major black shank resistance QTL, a trait that was successfully localized using the 14 recombinant lines.

Conclusions

This study was successful in enhancing our understanding of a tobacco introgression fragment with historically useful blue mold resistance characteristics. It was also successful in generating 14 recombinant lines in which the introgression fragment had been shortened, potentially laying the groundwork for the development of blue mold resistant lines with reduced yield drag. RNA-seq and recently available whole genome sequence information was used to find and map 26 contigs to the introgression fragment. Finally, analysis of the blue mold tolerant variety NC 2000 suggested that maximal blue mold

resistance is attained by combining the resistance QTL located within the *N. debneyi* introgression fragment with additional resistance QTL of *N. tabacum* origin.

BACKGROUND

Tobacco (*Nicotiana tabacum*) has long been an important model organism and valuable commercial crop. *N. tabacum* is an allotetraploid, possessing a large repetitive genome, a feature that can hinder genomic investigation. However, the emergence of streamlined next generation sequencing protocols and the recent availability of tobacco draft genomes and transcriptomes has made genomics-based investigations increasingly feasible (Sierro et al., 2014; Edwards et al., 2017). Many useful phenotypic and disease resistance traits in tobacco and other crops were originally identified in their respective wild relatives and later transferred into commercial breeding varieties (Lewis, 2005; Harlan, 1975). Upon the transfer of these traits, additional, often detrimental, linked genes are frequently included (Wernsman, 1999). Further, beneficial genes within the recipient plant can be displaced when a new segment is introgressed (Wernsman, 1999). These issues (the former commonly termed linkage drag) are often first addressed through breeding methods such as backcrossing, or crossing to a line whose properties compensate for the linked detrimental effects (Ganal and Tanksley, 1991). Targeted reduction of an introgression fragment can be accomplished though recombination. However, if the species from which

the fragment originated is distantly related to the host genome, recombination may only occur at a low or unobservable level (Alpert and Tanksley, 1996).

Tobacco blue mold, caused by *Peronospora tabacina* D.B. Adam (syn. *Peronospora hyoscyami* de Bary), is an obligate biotrophic oomycete that can infect tobacco throughout the world, mostly in warmer climates (Spring et al., 2013). In some years the disease has caused tens of millions of dollars in damage in North America alone (Wu et al., 2015). Although blue mold does not overwinter in northern climates, it is spread through transfer of infected plants or spores traveling in the wind (LaMondia and Aylor, 2001; Blanco et al., 2017). Symptoms are manifest as light brown to light blue fuzzy downy mold within necrotic leaf patches (Sukanya and Spring, 2013). Under conditions favoring high spore germination (high humidity and low direct sunlight), plants without native resistance are often killed when systemic infection occurs (Spring et al., 2013).

Fungicides such as metalaxyl, azoxystrobin, and dimethomorph have been used to control blue mold in the field, but can become expensive if used preemptively and are proving less effective as resistant blue mold strains arise (Derevnina et al., 2015**;** LaMondia, 2013). When available, the introduction or enhancement of host resistance often provides the least expensive, most environmentally friendly, and most effective means for control of plant diseases.

A number of chemicals found either within the cell or on the leaf surface of tobacco, including cembratriene diols (α-CBT-diols and β-CBT-diols), labdenediol, and sclareol have been shown to have a negative effect on germination of *P. tabacina* sporangiospores when

in concentrations found in some tobacco cultivars (Kennedy et al., 1992). Tobacco genes encoding β-Ionone, PR-1a, glucanase, gluthione synthetase, an EIL2 transcription factor and T-phylloplanin have also been experimentally shown to have inhibitory effects on infection of tobacco by *P. tabacina* (reviewed in Borras-Hidalgo et al., 2009).

Over 50 years ago, it was shown that the wild tobacco species *N. debneyi* and *N. goodspeedii* were resistant to infection by the blue mold pathogen. Wide crosses were successful in transferring portions of the DNA from wild species into cultivated tobacco that conferred some blue mold resistance, though not as complete as that found in the wild species (Clayton, 1967; Wark, 1963). The best characterized and most widely used source of the blue mold resistance in tobacco is derived from an introgression fragment originating from *N. debneyi* (Milla et al., 2005). In plants where the introgression fragment (referred to as *bmr*) is homozygous (such commercial varieties NC 2000 and NC 2002), there is an associated yield penalty. Resistance is reduced when the introgression fragment is heterozygous (e.g. hybrid variety KT206). Resistance provided by *bmr* has been shown to be slow acting, allowing *P. tabacina* to infect and ultimately leads to a hypersensitive response (HR) (Wu et al., 2015). Incorporation of the *bmr* introgression fragment can been tracked in tobacco from modern breeding programs using three dominant sequence characterized amplified region (SCAR) coupling phase markers expected to be closely linked (Milla et al., 2005; Julio et al., 2006). An additional repulsion phase marker for the introgression fragment has been developed based on a Bindler et al. (2011) marker designated PT61512,

and the *bmr* integration site has been localized to one end of chromosome seven (Wu et al., 2015).

In this study, we investigate the *bmr* introgression fragment and its relationship to two SCAR markers that have been reported to flank the gene(s) conferring blue mold resistance (Wu et al., 2015). RNA-seq was used to identify genes of *N. debneyi* origin that are located on the introgression fragment. Recombinant lines were developed and used in conjunction with markers derived from the RNA-seq data to generate a map of the region. Due to difficulties in conducting field experiments where blue mold pressure was sufficient to accurately phenotype the recombinant lines, fine mapping of the chromosomal region responsible for conferring enhanced resistance to blue mold was not successful. Unexpectedly, however, the *bmr* introgression fragment was found to replace a region of the *N. tabacum* genome possessing a major resistance QTL to the tobacco oomycete disease black shank (*Phytophthora parasitica* var. *Nicotiana*), which we were able to map within the context of the *N. debneyi*-derived introgression fragment.

RESULTS

Screening for recombinants

An initial objective of this project was to confirm that the *bmr* introgression fragment in NC 775 bmr/bmr was capable of active recombination with the corresponding *N. tabacum* chromosomal region. Due to either genetic distance or introgression occurring

randomly at a non-homologous genomic location, recombination may not occur with interspecific introgression fragments introduced from wild relatives (Alpert and Tanksley, 1996). In a previous study, the *bmr* introgression fragment was mapped using random amplified polymorphic DNA (RAPD) markers (Milla et al., 2005). Two of the RAPD markers, predicted to span a genetic distance of ~6 cM, and flank the region conferring blue mold resistance, were converted into more facile SCAR markers, designated SUBC180.251 and SOPR06.268 Milla et al., 2005). In our report, markers SUBC180.251 and SOPR06.268 are referred to as SCAR 251 and SCAR 268, respectively. The same SCAR markers were recently used by Wu et al. (2015) who estimated the genetic distance between SCAR 251 and SCAR 268 to be 3.3 cM.

Tobacco line NC 775 bmr/bmr was developed by backcrossing the burley variety NC 2000 to NC 775 as the recurrent parent for six generations (followed by self-pollination to fix the trait) while selecting for the *bmr* locus using the SCAR 251 and SCAR 268 markers. To generate recombinant lines for fine mapping, as well as validate the results of Milla et al. (2005) that suggested the introgression region was recombinationally active, crosses were made between the near isogenic lines NC 775 and NC 775 bmr/bmr. The BC_7F_1 progeny resulting from this cross were genotyped using SCAR 251 and SCAR 268. Recombinant individuals were recognized as those lacking of one of the flanking SCAR markers. A typical example of a genotyping gel is shown in Figure 1. The screening of approximately 2,700 BC_7F_1 individuals resulted in the recovery of 14 NC 775 bmr* recombinant individuals (the asterisk signifying the presence of a partial introgression fragment), eight lacking SCAR 268

and six lacking SCAR 251. Given that the SCAR markers are dominant, it would be fully expected that the great majority of individuals showing a pattern indicative of a recombination event would be heterozygous at the allele containing the SCAR marker, as homozygous individuals would only be observed if two independent recombination events had occurred in the region of interest. In order to obtain lines that were fixed for the partial introgression fragments, each of the original 14 recombinant plants was self-pollinated and the progeny were genotyped for the presence of the relevant SCAR marker. Individuals possessing the marker were test-crossed to a wild type tobacco plant, and the offspring of the test-crosses scored again with the SCAR marker to identify parental plants that were homozygous for the partial *N. debneyi* introgression fragment. This procedure was successful in developing fixed lines (NC 775 bmr*/bmr*) for all 14 original recombination events.

Sequencing and transcriptome generation

As an initial characterization of the *N. debneyi*-derived *bmr* introgression fragment found in NC 775 bmr/bmr, RNA-seq was performed on NC 775, NC 775 bmr/bmr and *N. debneyi* (accession TW 36 from the North Carolina State University *Nicotiana* germplasm collection). In addition to providing insights with respect to some of the specific *N. debneyi* genes present on the introgression fragment, a primary goal of the RNA-seq analysis was to develop markers that could be mapped using the 14 recombinant NC 775 bmr*/bmr* lines. Contigs that displayed high homology between NC 775 bmr/bmr and *N. debneyi*, and poor

NC 775 homology could be considered as introgression fragment candidates. RNA-seq analysis generated raw reads counts of 92,889,432 for NC 775, 49,565,076 for NC 775 bmr/bmr, and 62,584,146 for *N. debneyi*. Respectively, 93.5%, 76.7%, and 84.6% of the raw reads were retained and utilized after quality control and trimming.

Although published genome and transcriptome resources exist for some *Nicotiana* species, including high quality assemblies of *N. tabacum*, very limited genetic resources currently exist for *N. debneyi* (Sierro et al., 2014; Edwards et al., 2017). Nevertheless, we predicted that the genomic regions within the *N. debneyi*-derived introgression fragment would be evolutionarily distant enough from available *N. tabacum* resources that *de novo* transcriptome assembly would be required. For this study, we processed RNA-seq reads from NC 775, NC 775 bmr/bmr, and *N. debneyi* to assemble a transcriptome. This resulted in 216,398 total trinity transcripts with an N50 of 1,289.

Candidate contig list creation through raw alignment number.

Discovery of contigs unique to the introgression fragment is expected to be easier than when looking for genes that are part of a deletion (as was the situation in Chapter 2). This, in addition to the genetic distance between NC 775 and *N. debneyi* means that the number of false positive contigs predicted to be part of the introgression fragment should be relatively minimal. The goal is to identify list of contigs that align to a large number of individual processed reads from *N. debneyi* and NC 775 *bmr*/*bmr*, and none from NC 775.

Each of the three data sets were aligned with the transcriptome individually and alignment statistics collected.

For marker production, an exhaustive list of all contigs that may be legitimately located within the introgressed fragment is not necessary. Therefore, a list of candidates was created based on very stringent alignment criteria. A list of candidate contigs was created where each required >49 aligned processed RNA-seq reads from both *N. debneyi* and NC 775 bmr/bmr and <2 aligned NC 775 reads. Under this level of stringency, a legitimate introgression fragment contig would be removed from consideration if it had an ~100bp length of sequence displaying near perfect homology to another part of any of the three transcriptomes. Illumina sequencing technologies has an ~1% error rate (Goodwin et al., 2016). Errors of this nature are unlikely to lead to removal of positive contigs as a read would need to have near perfect homology to the introgression contig **and** be mis-reported by the error at exactly the non-homologous base pair(s). Furthermore, those contigs that do display high homology to other areas within the genome typically make poor marker candidates due to the difficulty in designing primers to uniquely amplify the region. Expression analysis was not performed in a classic way, as the discovery of a complete list of differentially expressed genes was not the primary goal. By conducting alignments using very high stringency parameters, we were able to quickly develop a candidate gene list with low probability for false positives. Also, it allowed for our candidate list to be expanded easily by simply imposing a smaller alignment threshold.

As a further constraint to the candidate contigs, all entries were BLAST searched against the *N. tabacum* K326 genome (Sierro 2014). K326 is a flue-cured variety and does not possess the introgressed *bmr* fragment, in addition to displaying low blue mold resistance overall (Rufty, 1989); therefore, legitimate candidate transcripts would not be expected to align well with this genome. Four contigs which showed exceptional alignment to K326 were removed as a result of imposing this criteria. The final list contained 152 contigs expected to be located within the *N. debneyi*-derived introgression fragment.

Contig validation and mapping

The list of 152 candidate contigs was annotated through BLAST searches using the NCBI databases, and is displayed in Table A3.1, Appendix. Three of these, 132209g2i2, 132209g2i3, and 139677g2i4 are annotated as having homology to putative late blight resistance proteins. A large percentage (78%) returned no annotation. In addition to the contigs listed in Table A3.2, Appendix, a less stringent contig list, corresponding to 396 more transcripts, was annotated to search for additional candidates that may be associated with disease resistance responses. The expanded list contained a higher percentage of contigs that could be annotated (many of which may be false positives), but none were found with annotations suggesting a potential role in disease resistance.

Primers were designed for 47 contigs from the list of 152, with an emphasis on contigs with annotations and limited repetitive sequence. PCR assays were conducted using genomic DNAs isolated from K326-va, NC 775, NC 775 bmr/bmr, and *N. debneyi*. Primers designed against 26 of the 47 transcripts (55%) amplified strong bands in NC 775 bmr/bmr

and *N. debneyi* that were absent in K326-va and NC 775, making them suitable for marker analysis (Table 1), Typical examples of the PCR validation experiments are shown in Figure 2. Primer sequences for the 26 *N. debneyi*-specific markers are shown in Table A3.1, Appendix.

Table 1: 26 PCR validated contigs and their annotations.

Contig ID	Marker ID	Annotation
129463g3i2	Rq-1	Alcohol dehydrogenase 1-like
132205g2i1	Rq-2	none
114942g1i1	Rq-3	Uncharacterized protein LOC104112417
125921g1i1	Rq-4	Uncharacterized protein LOC103439082, partial
136452g7i1	Rq-5	none
165520g1i1	Rq-6	none
142993g3i3	Rq-7	none
134479g1i5	Rq-8	none
137710g2i2	Rq-9	none
60182g2i1	Rq-10	none
139743g2i4	Rq-11	Uncharacterized protein LOC104085741
134994g1i7	Rq-12	none
1271g2i1	Rq-13	none
131019g1i3	Rq-14	none
134941g1i3	Rq-15	none
136458g1i2	Rq-16	none
137017g2i6	Rq-17	Nucleobase-ascorbate transporter 4-like
136662g7i1	Rq-18	none
139654g6i1	Rq-19	none
121166g2i3	Rq-20	none
126474g1i2	Rq-21	none
126444g2i2	Rq-22	none
140362g5i1	Rq-23	Uncharacterized protein LOC108945409
122515g3i1	Rq-24	none
126184g2i4	Rq-25	Uncharacterized protein LOC109216090
132209g2i3	Rq-26	Putative late blight resistance protein

PCR reactions using the primers for the 26 verified contigs were run for each of the 14 recombinant NC 775 bmr*/bmr* lines. By assaying for the presence or absence of each of the 26 contigs across all 14 recombinant lines, we were able to establish the relative order of the contigs across the original *bmr* introgression fragment, and roughly determine the extent of *N. debneyi* DNA remaining in each NC 775 bmr*/bmr* lines, as depicted in Figure 3. A map of this region, with the number of recombination events observed between each marker is shown in Figure 4.

Investigation of genetic synteny with other *Nicotiana* genomes

Although no significant genome sequence information currently exists for *N. debneyi* or any tobacco line containing the *bmr* introgression fragment, current published tobacco genome data may still be useful in characterizing this region if syntenic relationships can be established. Although *N. debneyi* has not been represented in any of the recent studies of relatedness among members of the *Nicotiana* genus (Fricano et al., 2012; Dadras et al., 2014), it is possible that orthologous stretches of DNA exist in one of the sequenced tobacco species and that this region could be used to predict additional genes that would likely be located within the introgressed fragment.

The 26 contigs used to create the marker map were searched with BLAST using a low quality threshold cutoff against four *Nicotiana* genomes, *N. tabacum* K326 (Edwards et al., 2017); *N. attenuata* and *N. obtusifolia* (Xu et al., 2017); and *N. benthamiana* (Bombarely et

al., 2012). The cutoff for what could represent a useful stretch of DNA was three transcripts matching to the same scaffold or a similar chromosomal region for one of the tobacco species. Unfortunately, when two of the contigs that share high homology to a large number of genes in these genomes were excluded from the analysis (136452g7i1 and 136662g7i1), the three-match threshold could not be met.

Field trials of parental *bmr* materials

Obtaining accurate phenotypic data on genetic-based resistance to blue mold has historically been difficult. Due to the sporadic, unpredictable nature of blue mold infestations, there are no disease nurseries in the U.S. where materials can reliably be tested for their response to the pathogen. Blue mold does tend to be more prevalent in more southerly, tropical climates such as certain regions of Mexico and the Caribbean. In the 2014/15 growing season, a field test for blue mold resistance was conducted in the Dominican Republic. This represented the first evaluation of the NC 775 bmr/bmr line. In addition to NC 775 bmr/bmr, the *bmr* introgression fragment from *N. debneyi* was similarly backcrossed into burley cultivar NC 645 (6 backcross generations, using SCAR 251 and SCAR 268 to select for the *bmr* locus). The resulting near-isogenic line, NC 645 bmr/bmr was also included in the Dominican Republic field trial. Damage attributed to blue mold was assessed using a 1 - 5 point scale established by a study panel commissioned by the international tobacco organization CORESTA, with 1 being no signs of infection to 5 being near dead. As shown in Table 2, the average ratings for disease severity was higher for lines NC 775 and

NC 645 than their *bmr*-containing counterparts. What was unexpected, however, was that NC 2000 was substantially more resistant to blue mold than either NC 775 bmr/bmr or NC 645 bmr/bmr, despite being the source of the *bmr* introgression fragment for both of these lines.

Table 2. Blue mold disease ratings of lines grown in the Dominican Republic in 2014.

Genotype	Average
NC 2000	2.44*
NC 775 bmr/bmr	3.62
NC 645 bmr/bmr	3.63
NC 775	4.05
NC 645	4.67

*ratings based on a 1 – 5 severity scale established by a CORESTA study panel; 5 = greatest disease damage

NC 2000 sequencing and analysis

The results of the blue mold field trial showed that NC 2000 displays substantially greater resistance to blue mold than was observed in two other burley lines possessing the *bmr* introgression fragment. In both cases the *bmr* region was transferred to the new parental lines through many generations of backcrossing using the SCAR 251 and SCAR 268 markers to select of the introgression region. There are two possible explanations for the increased resistance in NC 2000: (1) NC 2000 has additional endogenous resistance QTLs of *N. tabacum* origin that were lost during the backcrossing of the *N. debneyi*-derived

chromosomal region into the new recurrent parents; or (2) some of the genes of *N. debneyi* origin that contribute toward the blue mold resistance phenotype are located outside of the region encompassed by SCAR 251 and SCAR 268, and were lost during backcrossing via recombination. If the latter scenario is true, then one should be able to find additional sequences of *N. debneyi* origin within NC 2000 that are not found in NC 775 bmr/bmr (or NC 645 bmr/bmr).

To test whether the NC 2000 genome encodes *N. debneyi*-derived genes that are not found in NC 775 bmr/bmr, we performed RNA-seq on NC 2000. Illumina-based sequencing of NC 2000 yielded 66,652,832 processed reads. In order to identify relevant contigs, two alignment strategies similar to the ones we conducted previously were employed. First, NC 2000 reads were aligned to our original transcriptome generated using NC 775, NC 775 bmr/bmr and *N. debneyi* reads. Contigs with a high number of alignments in both NC 2000 and *N. debneyi* but few in NC 775 and NC 775 bmr/bmr were considered as viable candidates. As transcriptome assemblies can vary based on read count and a diversity of inputs, a second list was created using a transcriptome assembled solely using NC 2000 processed reads. From this, contigs with a high number of aligning NC 2000 and *N. debneyi* reads and low levels of aligning NC 775, NC 775 bmr/bmr, and K326 reads (the K326 reads used [SRR955772] were those deposited by Edwards et al., 2017). Top candidates from both lists were selected and primers were synthesized for PCR validation. Sixty-four primer pairs representing the top 40 contigs were tested with no amplification patterns observed that would support the proposal that there is *N. debneyi* DNA present in NC 2000 that had been

lost when creating NC 775 bmr/bmr. Interestingly, as shown in Figure 5, in a couple of cases amplification products were observed that were either unique to NC 2000, or were shared with the flue-cured variety K326 and not with either of the other two burley lines tested (NC 775 and NC 775 bmr/bmr). These results support the model that one or more genes located on the *N. debneyi* introgression fragment contribute to the blue mold resistance phenotype of NC 2000, but that other non-*N. debneyi* resistance QTL are also important to confer maximal disease resistance.

Disease resistance trials for NC 775 bmr*/bmr* lines

A seminal goal of this project was to develop tobacco lines that retain the blue mold resistance-conferring portion of the *N. debneyi*-derived introgression fragment, while eliminating as much of the extraneous *N. debneyi* sequences as possible in an attempt to reduce the associated yield drag. This requires field analysis of the various fixed NC 775 bmr*/bmr* lines in an environment conducive for blue mold infestation, which as mentioned above, can be problematic with this disease. An interesting, unexpected observation that became apparent during the course of this project, however, was that the parental line NC 775 bmr/bmr, as well as other backcross-generated line NC 645 bmr/bmr, were both exceptionally susceptible to the disease black shank, caused by the soil-borne fungal pathogen *Phytophthora nicotianae*. Unlike blue mold, field evaluation for black shank resistance can be readily tested in black shank nurseries located in the U.S. that consistently harbor high levels of inoculum for the disease.

To formally test the black shank resistance phenotypes of NC 775, NC 775 bmr/bmr, and the various NC 775 bmr*/bmr* lines developed in this project, each of these lines were grown in a black shank nursery in Rocky Mount, North Carolina and evaluated for black shank susceptibility. As shown in Figure 6, the Area Under Disease Progress Curve (AUDPC) for NC 775 bmr/bmr (7.8) was far greater than its near-isogenic partner line NC 775 (1.7), confirming prior informal observations suggesting that the two lines differed substantially in their response to black shank infection (data produced by Dr. Justin Ma under the direction of Dr. Ramsey Lewis, NCSU). Interestingly, the recombinant NC 775 bmr*/bmr* lines included in this trial (fixed lines had been developed for 11 out of the 14 original recombination events that the time of this study) could be grouped into one of two groups, one showing a susceptibility response similar to NC 775 bmr/bmr, and the other group displaying a resistance phenotype similar to NC 775 (Fig. 6). By aligning the black shank resistance phenotypes of individual NC 775 bmr*/bmr* lines (Fig. 6) to the location and proportion of remaining *N. debneyi*-derived sequences in that line (Fig. 3), the data perfectly fit a model whereby lines containing *N. debneyi* sequences from the markers SCAR-251 up to (and possibly including) Rq1 retain the major black shank resistance QTL found in NC 775, while lines possessing the *N. debneyi* sequences containing SCAR-268 (including the extent of unknown *N. debneyi* DNA preceding this marker), up to (and possibly including) Rq1 (Fig. 4).

In contrast to the field evaluation for black shank, finding a suitable and effective environment for blue mold testing proved to be difficult. The NC 775 bmr*/bmr* lines

together with the NC 775, NC 775 bmr/bmr and NC 2000 controls were planted in the Dominican Republic in the Fall of 2016 (in accordance to their growing season). In contrast to the trial planted in the Fall of 2014 (at which time the fixed NC 775 bmr*/bmr* lines had yet to be developed), minimal blue mold was evident in the field. Furthermore, a high incidence of blank shank (or phenotypically similar soil-borne disease, as precise pathogen identification could not be conducted due to laws prohibiting international transport of pathogens without a permit) caused widespread destruction on all lines shown as being black shank susceptible in Figure 6. A representative picture of a portion of this field is shown in Figure 7. Thus, even if there had been sufficient blue mold pressure to distinguish the resistance phenotypes of the recombinant lines, interpretable data with respect to blue mold would still not have been possible due to the high incidence of plant death and disease caused by black shank in this field.

DISCUSSION

In this study we conducted a molecular genetic characterization of the *N. debneyi*-derived *bmr* introgression fragment that has been commercially utilized as a source of blue mold resistance in commercial tobacco varieties such as NC 2000 and KT 206. Although the overall size and extent of the introgression fragment is unknown, a recent mapping study has placed it on *N. tabacum* linkage group 7 (Wu et al., 2015). This same report mapped the blue mold conferring gene(s) within this region to be localized within a 3.3 cM region

flanked by the SCAR 251 and SCAR 268 markers developed by Milla et al. (2005)(in the original study by Milla et al., the genetic distance between these markers was estimated to be 6.1 cM.) Using RNA-seq, we were able to create a list of 152 contigs generated using very stringent alignment criteria as a means of identifying genes with a high probability of originating from *N. debneyi*, and thus for being located on the introgression fragment. In addition to providing insights with respect to the gene composition of this region, these results also laid the groundwork for marker development to facilitate the fine mapping of this region.

When developing markers for mapping studies, PCR primers that generate robust, unambiguous amplification products with a limited number of bands are preferred. This is often more difficult to achieve in species with highly repetitive genomes such as tobacco, requiring the testing of many candidate primers. In this particular study it was beneficial that the target introgression fragment came from a wild tobacco relative so as to increase the divergence to analogous regions within the *N. tabacum* genome. Under this premise we believed that a candidate contig selection methodology with very high stringency would yield a high percentage of legitimate gene candidates. Forty-seven of the 152 contigs from our original candidate list were PCR tested in *bmr* versus non-*bmr* lines, with 26 giving a pattern matching what would be expected for *bmr* introgression fragment inclusion (Fig. 2 and Table 1). Although 21 of the 47 contigs tested were not validated, this does not mean that they were false positives, as a variety of reasons were responsible for this, such as the failure to design primers that would yield unambiguous, robust amplification products, the

presence of highly homologous sequences elsewhere in the genome that complicated the ability to generate specific primers, or the potential presence of introns interfering with primer design based on the transcriptome.

In an attempt to find *Nicotiana* scaffolds that could be used to place some of the *N. debneyi* contigs and take advantage of syntenic relationships to predict additional genes on the introgression fragment, available draft genomes of four *Nicotiana* species were aligned with the 26 validated contigs. In particular, one of the investigated species, *N. obtusifolia,* has been shown to display resistance interactions with *P. tabacina* (blue mold) through an HR response involving the partially dominant Rpt1 gene (Heist et al., 2004). While no candidate scaffolds were found among the four species to enable this, there remains the possibility that if additional *bmr* contigs are validated, scaffolds of this nature may be identified. Furthermore, in the draft genomes of the tobacco species investigated, scaffolds where one of the 26 contigs did show substantial alignment were often short and did not contain any additional genes. If improved genome drafts are released in the future, repeating this line of investigation may prove fruitful.

Twenty-six PCR verified markers were mapped to the 14 recombinant NC 775 bmr*/bmr* lines (Figure 3). These markers enabled the placement of these lines into nine distinct groups. The markers generated in this study occur as singles with the notable exception of a large clustering of 17 that grouped together with SCAR-268. One possible explanation is that the chromosomal region around SCAR-268 may be particularly gene rich. More likely, however, is that many if not all of these 17 contigs fall on the introgression

fragment outside the region flanked by SCAR-268 and SCAR-251. Of necessity, the selection of recombinants from the NC 755 bmr BC_7F_1 materials resulted in individuals possessing either SCAR-251 or SCAR-268. The extent of *N. debneyi* sequences on either side of these markers is unknown, and could be substantial. The longer the introgression fragment extends beyond either SCAR marker, the greater the likelihood of identifying contigs that will "co-localize" with the terminal SCAR marker due to the experimental design. Assuming there is a fairly uniform spacing of genes across the introgression fragment, the results shown in Fig. 4 would predict that the extent of the *N. debneyi*-derived introgression fragment beyond the SCAR-268 marker (with 17 co-localizing contigs) is far greater than that which extends beyond SCAR-251 (with 3 co-localizing contigs).

In addition to serving as a source for markers for fine mapping, the contigs identified using RNA-seq also have the potential of being directly involved in the resistance phenotype. An uncharacteristically low number of contigs identified by RNA-seq, however, could be annotated (see Table A3.2, Appendix). Among the annotated contigs, however, are three that display homology to putative late blight resistance proteins, one of which was validated and converted into marker Rq-26 (132209g2i3). The three contigs encoding putative late blight resistance protein were the only candidates annotated as having any association with a disease resistance response, therefore further investigation of these may be warranted. Even when the alignment stringency was relaxed to expand the candidate list to include an additional 396 contigs, no more disease-associated annotations were found.

In the successful field test conducted in the Dominican Republic during the 2014/15 growing season that included the parental lines use this study, it was shown that NC 2000, the source of *bmr* in NC 775 bmr/bmr and NC 645 bmr/bmr (another burley line produced in this same manner), provided superior blue mold protection than NC 775 bmr/bmr or NC 645 bmr/bmr. This finding prompted RNA-seq analysis to determine whether additional *N. debneyi* sequences exist in NC 2000 that had been lost through recombination during the development of NC 775 bmr/bmr (and likely NC 645 bmr/bmr as well). This investigation failed to find any additional contigs of *N. debneyi* origin in NC 2000, supporting the concept that there are unique blue mold resistance QTLs within NC 2000 not are present in NC 775 and were lost in the backcrossing scheme that lead to the production of NC 775 bmr/bmr. Thus, in order to take advantage of the full blue mold resistance potential of NC 2000, it would be advantageous to develop markers against these putative additional QTLs, though such an effort would likely be hindered by the inherent difficulties associated with field evaluation of blue mold.

The linkage drag found in plants homozygous for the *bmr* introgression fragment is likely unrelated to the specific gene(s) conferring blue mold resistance. The 14 recombinant NC 775 bmr*/bmr* lines developed in this study represent an array of materials possessing various portions of the original introgression fragment. Among these may be lines that retain the blue mold resistance trait and display reduced yield drag due to the shortening of the region of chromosome 7 that is represented by *N. debneyi* DNA. In order to test this, it will first be necessary to identify the specific recombinant lines that contain the resistance

bmr QTL, which can only be accomplished through field testing. Unfortunately, blue mold is among the most difficult diseases to assay in the field, with a high degree of variability and unpredictability. Our attempt to phenotype the NC 775 bmr*bmr* line in the Dominican Republic during the 2016/17 field season failed, in part due to insufficient blue mold pressure.

The other primary reason that the 2016/17 field trial proved to be problematic was due to the unexpected phenomenon that tobacco plants possessing the *N. debneyi* introgression fragment are especially susceptible to black shank, leading to the near or total destruction of all recombinant lines that had been independently shown to be susceptible to black shank (Figs. 6 and 7). The susceptibility of the NC 775 bmr/bmr parental line and several of the NC 775 bmr*/bmr* lines to black shank may be attributable to either the presence of undesirable susceptibility-associated genes from *N. debneyi*, or the absence of an endogenous *N. tabacum* black shank resistance QTL due to its replacement by the introgression fragment. The latter explanation is greatly supported by the recent mapping of a major black shank resistance QTL to chromosome 7 of tobacco (Ma J, 2017; Vontimitta and Lewis, 2012). Further investigation of the relationship between the black shank resistance QTL on chromosome 7 and the *bmr* resistance QTL is warranted. Of particular interest will be determining whether any of the NC 775 bmr*/bmr* lines developed in this study contain both the blue mold and black shank resistance QTLs, as these would be the most valuable materials from a variety breeding perspective.

MATERIALS AND METHODS

Sequencing and alignment

RNA for all sequencing projects was isolated using the RNeasy Plant Mini Kit (Qiagen). RNA quality verification was performed using a 2100 Bioanalyzer (Agilent Technologies). NEBNext Ultra Directional Library Prep Kit with Multiplex Oligos (these are comparable to TruSeq oligos) were used for RNA-seq prep of NC775, NC775 bmr/bmr, and TW36 N. debneyi. The three multiplexed samples were run on two lanes of an Illumina HiSeq chip using 100bp paired-end sequencing. NC 2000 RNA-seq was performed using Nextseq 150bp SE (Illumina Inc. San Diego, CA, USA). For false positive control, the K326 RNA-seq reads deposited in GenBank by Sierro et al. 2014 (SRR955772) were also used. Trimmomatic (Bolger et al., 2014) with settings, ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 HEADCROP:12 TRAILING:20 SLIDINGWINDOW:3:21 MINLEN:70, was used for initial read processing. NC 2000 used NIMLEN:90 and SE adaptor trimming. A NC775, NC775 bmr/bmr, and TW36 N. debneyi read pooled transcriptome was assembled using Trinity (v2.2.0) (Grabherr et al., 2011; Hass et al., 2013). Alignment of processed RNA-seq reads to the de novo transcriptome was performed using BWA (v0.7.12) (Li, 2013). Alignment data was collected with Idxstats (Samtools).

PCR analysis

PCR amplification of contigs was performed with primers listed in Table A3.1, Appendix. *Taq* DNA Polymerase (New England BioLabs, Ipswich, MA) was used with thermal cycling conditions consisting of an initial denaturing of 95°C for 30s; 32 cycles of denaturation at 95°C for 10s, annealing at 60°C for 20s, and extension at 68°C for 1m and a final cycle of extension at 68°C for 5m.

REFERENCES

Blanco M, Carbone I, Ristaino J. Population structure and migration of the Tobacco Blue Mold Pathogen, *Peronospora tabacina* into North America and Europe. Mol. Ecol. 2017;In press.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–2120.

Grabherr MG, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnol. 2011;29:644–652.

Haas BJ, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494–1512.

Sierro N, et al. The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun. 2014;5:3833.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;arXiv preprint arXiv:1303.3997.

Ganal MW, Tanksley SD. Recombination around the Tm2a and Mi resistance genes in different crosses of Lycopcrsicon pimpinelliforlium. Theor. Appl. Genet. 1991;92:935-951.

Alpert KB, Tanksley SD. High-resolution mapping and isolation of a yeast artificial chromosome contig containing fw2.2-a major fruit weight quantitative trait locus in tomato. PNAS. 1996;93:15503-15507.

Fricano A, Bakaher N, Corvo MD, Piffanelli P, Donini P, et al. Molecular diversity, population structure, and linkage disequilibrium in a worldwide collection of tobacco (*Nicotiana tabacum L.*) germplasm. BMC Genet. 2012;13:18.10.1186/1471-2156-13-18.

Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller A, Martin GB. A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research. Molecular Plant-Microbe interactions. 2012;25:1523-1530.

Xu S, et al. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. PNAS. 2017;114:6133-6138.

Dadras AR, Sabouri H, Nejad GM, Sabouri A, Shoai-Deylami M. Association analysis, genetic diversity and structure analysis of tobacco based on AFLP markers. Mol Biol Rep. 2014;41:3317-3329.

Wu X, Li D, Bao Y, Zaitlin D, Miller R, Yang S. Genetic dissection of disease resistance to the blue mold pathogen, *Peronospora tabacina*, in tobacco. Agronomy. 2015;4:555-568.

Sukanya SL, Spring O. Influence of temperature and ultra-violet light on viability and infectivity of *Peronospora tabacina* sporangia. Crop Protection. 2013;51:14-18

Spring O, Hammer TR, Zipper R, Billenkamp N. Population dynamics in tobacco blue mold incidences as a consequence of pathogen control and virulence performance of *Peronospora tabacina* phenotypes. Crop Prot. 2013;45:76-82.

Derevnina L, Chin-Wo-Reyes S, Martin F, Wood K, Froenicke L, Spring O, Michelmore R. Genome sequence and architecture of the tobacco downy mildew pathogen *Peronospora tabacina*. Mol Plant-Microbe Interactions. 2015;11:1198-1215.

LaMondia JA. Reduced sensitivity of *Peronospora tabacina*, causal agent of tobacco blue mold, to Dimethomorph fungicide in Connecticut. Tob Sci. 2013;50:19-24.

LaMondia JA, Aylor DE. Epidemiology and management of a periodically introduced pathogen. Biological Invasions. 2001;3:273-282.

Kennedy BS, Nielsen MT, Severson RF, Sisson VA, Stephenson MK, Jackson DM. Leaf surface chemicals from *Nicotiana* affecting germination of *Peronospora tabacina* (adam) sporangia. Journal of Chemical Ecology. 1992;9:1467-1479.

Borras-Hidalgo O, Thomma BPHJ, Silva Y, Chacon O, Pujol M. Tobacco blue mould disease caused by *Peronospora hyoscyami* f. sp. *tabacina*. Mol Plant Path. 2009;11:13-18.

Clayton EE. The transfer of blue mold resistance from *Nicotiana debneyi*. Part III. Development of a blue mold resistant cigar wrapper variety. Tob Sci. 1967;11:107-110.

Wark DC. *Nicotiana* species as sources of resistance to blue mold (*Peronospora tabacina* Adam) for cultivated tobacco. In Proceedings of the 3rd World Tobacco Science Congress, Salisbury, Southern Rhodesia, 18–26 February 1963, Tobacco Research Board: Harare, Zimbabwe. 1963; pp. 252-259.

Rufty RC. Genetics of host resistance to tobacco blue mold. In Blue Mold of Tobacco; McKean, W.E., Ed.; American Phytopathological Society: St. Paul, MN, USA. 1989; pp. 141-164.

Milla SR, Levin JS, Lewis RS, Fufty RC. RAPD and SCAR markers linked to an introgressed gene conditioning resistance to *Peronospora tabacina* D.B. Adam. in tobacco. Crop Sci. 2005;45:2346-2354.

Julio E, Verrier JL, Dorlhac de Borne F. Development of SCAR markers linked to three disease resistances based on AFLP within *Nicotiana tabacum* L. Theor Appl Genet. 2006;112:335-346.

Bindler G, Plieske J, Bakaher N, Gunduz I, Ivanov N, Van der Hoeven R, Ganal M, Donini P. A high density genetic map of tobacco (Nicotiana tabacum L.) obtained from large scale microsatellite marker development. Theor Appl Genet. 2011;123:219–230.

Edwards KD, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans AD, Bombarely A, Allen F, Hurst R, White B, Kernodle SP, Bromley JR, Sanchez-Tamburrino JP, Lewis RS, Mueller LA.. A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. BMC Genomics. 2017;18:448.

Lewis RS. Transfer of resistance to potato virus Y (PVY) from Nicotiana africana to Nicotiana tabacum: possible influence of tissue culture on the rate of introgression. Theor Appl Genet. 2005;110:678-687.

Harlan JR. Genetic resources in wild relatives of crops. Crop Science. 1975;3:329-333.

Wernsman EA. An overview of tobacco breeding - past, present and future. Rec Adv Tob. 1999;25:5-35.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Gevet. 2016;17:333-51.

Sierro N, et al. Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. Genome Biol. 2013;14:R60.

Heist EP, Zaitlin D, Funnell DL, Nesmith WC, Schardl CL. Necrotic lesion resistance induced by Peronospora tabacina on Leaves of Nicotiana obtusifolia. Phytopathology. 2004;94:1178–1188.

Murphy JP, Cox TS, Rufty RC, Rodgers DM. A representation of the pedigree relationships among flue-cured tobacco cultivars. Tob Sci. 1987;31:70-75.

Ma J. The fine mapping of two black shank resistance loci and identification of a hybrid lethality gene in tobacco. PhD Dissertation NCSU. 2017.

Vontimitta V, Lewis RS. Mapping of quantitative trait loci affecting resistance to *Phytophthora nicotianae* in tobacco (*Nicotiana tabacum* L.) line Beinhart-1000. Mol Breed. 2012;29:89–98.

FIGURES

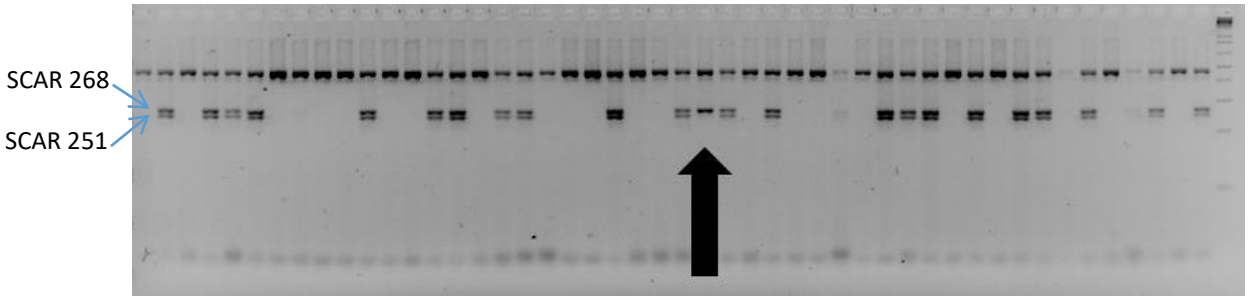
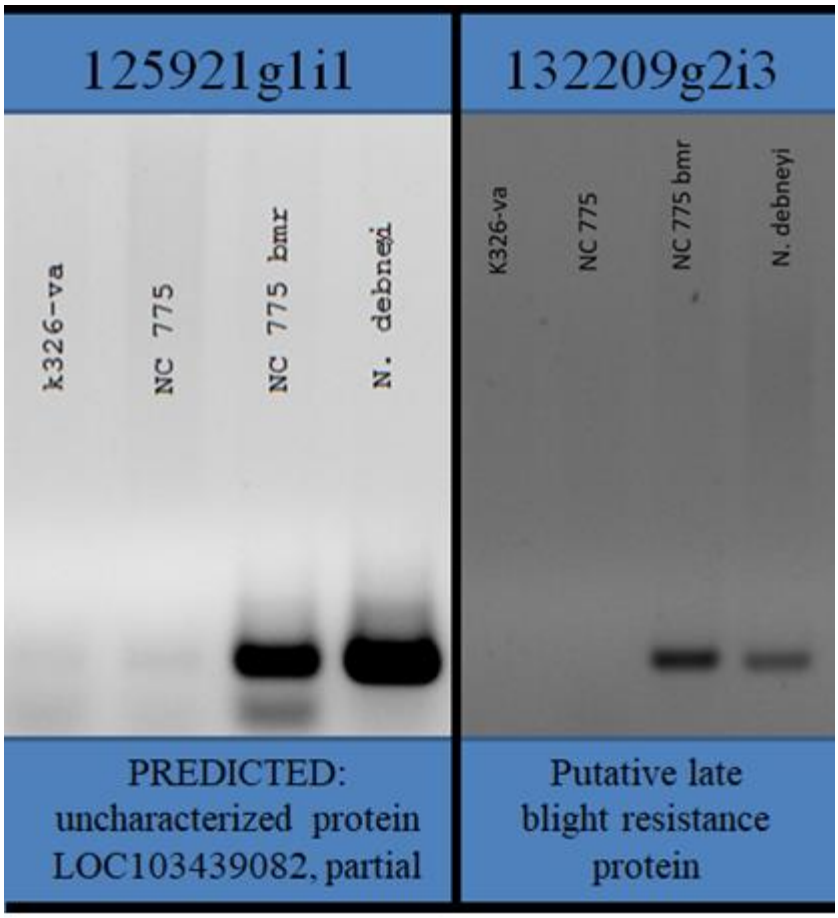


Figure 1.

SCAR marker analysis of BC_7F_1 progeny of a cross between NC 775 and NC 775 bmr/bmr.

Marker SCAR 251 produces an amplification product of 251bp; SCAR 268 generates a 268bp product. An example of an individual possessing a recombination event between the two markers is indicated by the large, black arrow.



F**igure 2.**

PCR verification for two contigs predicted to be included in the *N. debneyi*-derived *bmr* introgression fragment and not present in K326-va or NC 775. Annotations are shown at the bottom of each gel picture.

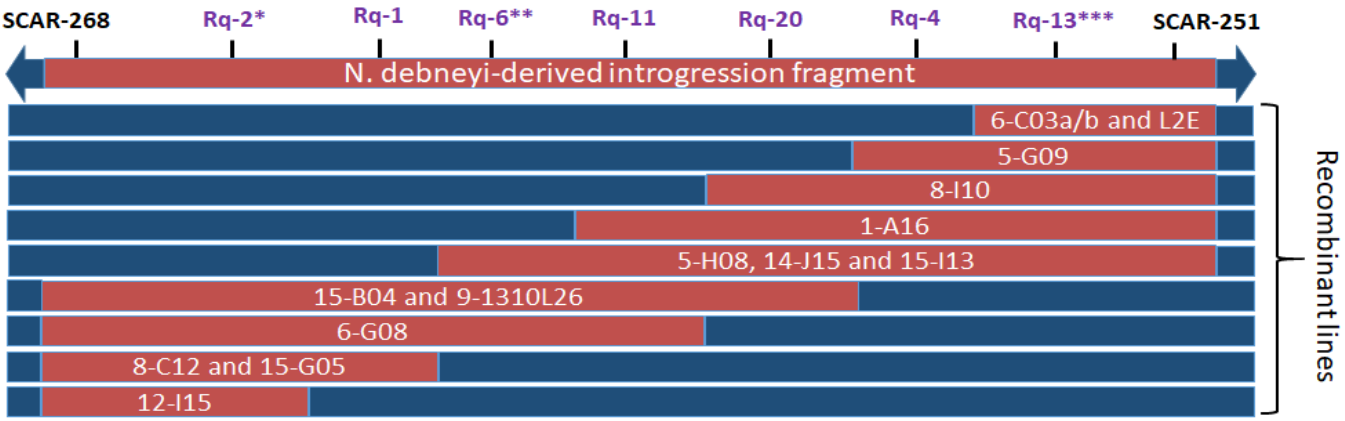


Figure 3.

Diagram showing the presence or absence of the 26 *N. debneyi*-specific contigs across all 14 NC 775 bmr*/bmr* lines. Spacing between markers is arbitrary and not to scale. *Sixteen additional markers map to this position (Rq-3, Rq-5, Rq-7, Rq-8, Rq-9, Rq-10, Rq-12, Rq-14, Rq-15, Rq-17, Rq-18, Rq-19, Rq-21, Rq-22, Rq-25, Rq-26); **one additional marker is represented at this position (Rq-16); ***two additional markers are represented at this position (Rq-23, Rq-24). Red background represents the *N. debneyi*-derived introgression fragment within NC 775 bmr/bmr and each of the recombinant lines. Blue represents genomic DNA of *N. tabacum* origin. Names of the 14 recombinant lines are labeled in white.

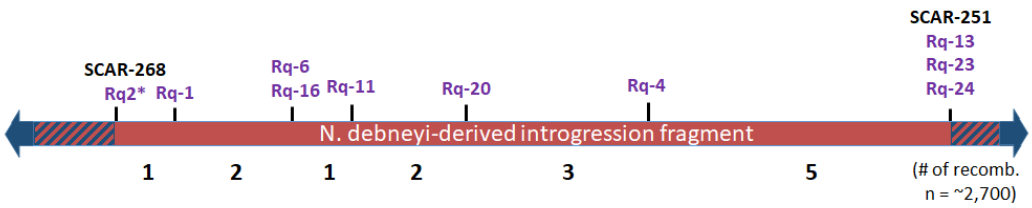


Figure 4.

Map of the *bmr* introgression fragment between markers SCAR-268 and SCAR-251.

Markers corresponding to the *N. debneyi*-derived contigs are indicated in purple type. *Sixteen additional markers map to this position (Rq-3, Rq-5, Rq-7, Rq-8, Rq-9, Rq-10, Rq-12, Rq-14, Rq-15, Rq-17, Rq-18, Rq-19, Rq-21, Rq-22, Rq-25, Rq-26). The number of recombination events observed between the various markers after genotyping approximately 2,700 BC_7F_1 individuals is shown beneath the diagram. *N. debneyi* sequences are shown in red, *N. tabacum* chromosomal DNA is depicted in blue, and the hatched regions indicate that the extent of the *N. debneyi*-derived fragment located outside the two SCAR markers is unknown.

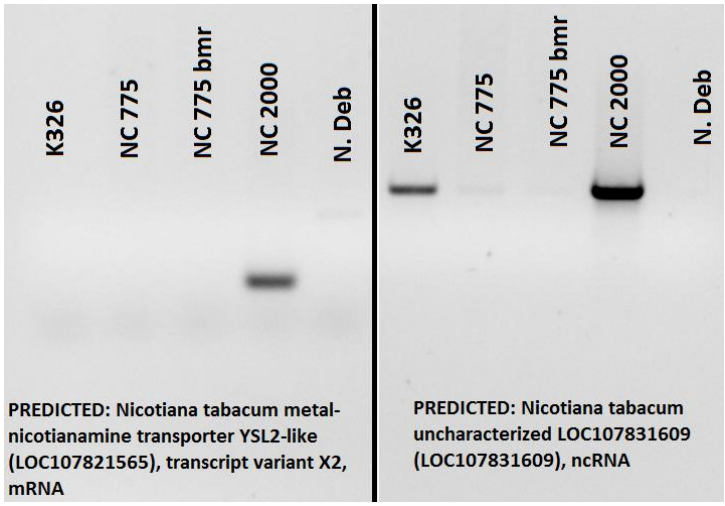


Figure 5.

Examples of candidate contigs that uniquely yielded clear amplification products only in NC 2000 (left panel) or NC 2000 and K326 (right panel).

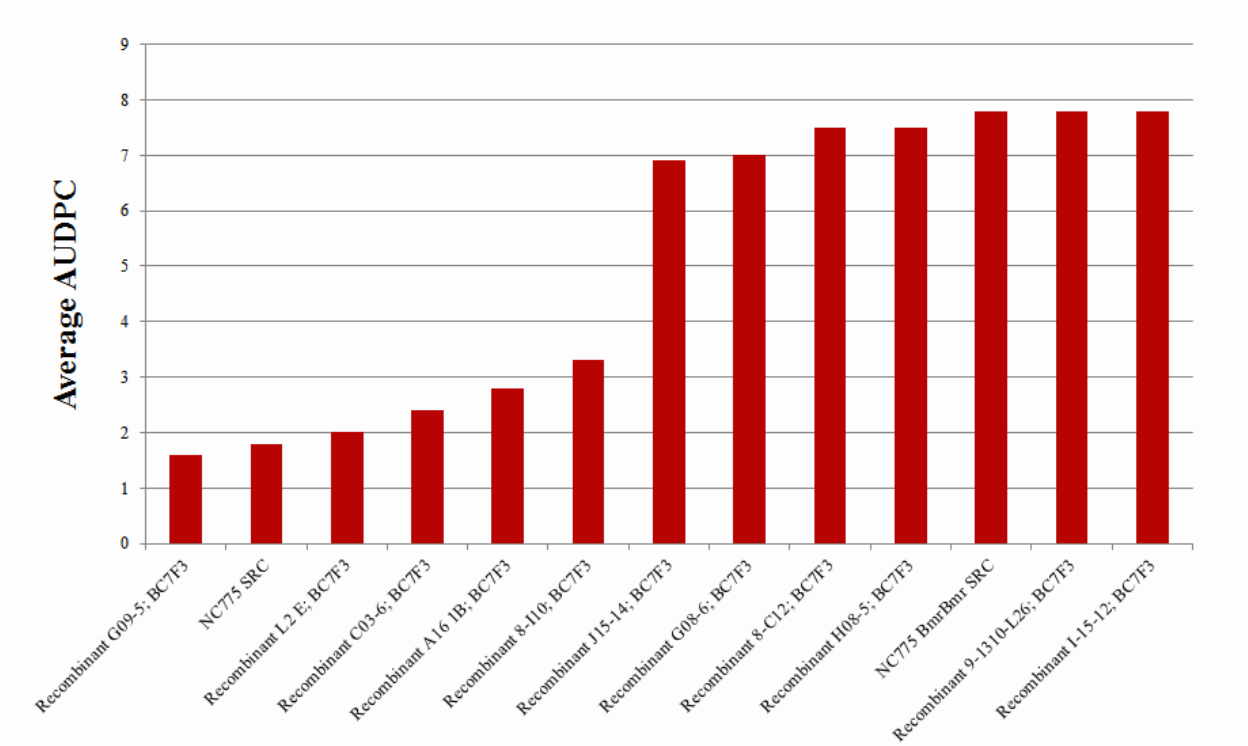


Figure 6.

The *bmr* introgression fragment masks a major black shank resistance QTL in NC 775.

Plants were grown in a black shank nursery in Rocky Mount, North Carolina. Susceptibility to black shack was scored by measuring Area Under Disease Progress Curve (AUDPC), using a scale of 0 through 10. (work performed by Dr. Ramsey Lewis of NCSU)

Figure 7.

Representative portion of the field trial of NC 775 bmr*/bmr* lines conducted in the Dominican Republic during the 2016/17 season. The three rows in front with small, missing (dead) plants represented lines that had been shown to be black shank susceptible in an independent trial conducted in the U.S. (see Fig. 6). Only minimal incidence of blue mold was observed in this field.

APPENDICES

APPENDIX

Table A2.1 Primer sequences

ID	Sequence
1441_3_F	ACATCAGGGAAATTGCAAGC
1441_3_R	TGCGTATGCATCTTCAAAGC
1441_2_F	CTAGTCCTCCTCCCCGTCTT
1441_2_R	AGTGTTGGGTCCATCTTTCGA
3331_1_F	CCATGTTTTAGGGGCTTTCA
3331_1_R	TTGGACATGTCGCACCTAGA
1441_2_4_F	ATTTGAAGGAGAGCCCTGGA
1441_2_4_R	AAACCTCGGGTCAATTCGAT
1441_2_5_F	CAGCGTTTGCAAAATCATTG
1441_2_5_R	AGCATATCATGCCGGAGTTC
1441_g1_F	TCGTGGAATCAGTTCTGCTG
1441_g1_R	CAGGAACCCGATACTTCACC
1441_g4_F	CCCTGACCTTGAGGACAAAA
1441_g4_R	GTACATGCACCCATGCTCTC
1441_g2_F	TGGCCAAATCAAGTGCATTA
1441_g2_R	AAGCAAGCTTGTTGACATGA
1441_g3_F	CGAACCCAATATGGTTGTCC
1441_g3_R	TTGCAAGAGCACAAACCTTG
23466c1g2i1F	TCTAAGGCTGCTGCTCCAAT
23466c1g2i1R	TTGACCCAAAGACCTGTTCC
24507c0g1i1F	CTCGCGAATTTTGTCCCTAA
24507c0g1i1R	CACTCAACAAATGAAGCAACG
22534c0g3i1F	TGCAGCTTTTATCGTGTTCG
22534c0g3i1R	TTCTCGCTCCTCCAACCTTA
35943c0g2i1F	CTGGCTTTTTGTGTCGTGAA
35943c0g2i1R	CACCACTCTCACACGCACAT
23466c1g2i2F	GAAAAGGTGCAAAACCAGGA
23466c1g2i2R	GACCCAAAGACCTGTTCCAA
42740c0g1i1F	TAGTGGCCTCGATAGGAACG
42740c0g1i1R	TGCCTCACGACATTCTTTTG
25436c0g1i2F	GTGGCAGTTTCAAGCCAGAT

Table A2.1 continued

25436c0g1i2R	ACGCCGAGTAATTCCACATC
86871c0g1i1F	CCAATCGGCTAAACCAAGAA
86871c0g1i1R	GCCGAACTTGCTGGAAGTAG
20225c0g1i1F	CCGACTTGAAACACCCATTT
20225c0g1i1R	ATTTGGCATTGGTCTTGGAG
18187c0g2i2F	CACCAACAGCTGGTCAAACA
18187c0g2i2R	GAAAGCGGATCAGGAATTTG
16240c0g1i1F	CGGTACCGTTGTTTCATGTG
16240c0g1i1R	ACTACCAACCGCTACCAACG
28570c0g1i1F	GAGGGAGGAATTTTGCATCA
28570c0g1i1R	GCCCTGCATCTCTTGAAATC
85964c0g1i1F	CCTCCCCTTCTTGTTCCTTC
85964c0g1i1R	ACTGTCACCCGTGATGTCAA
26611c0g1i2F	GCCACCCTTAATTCCTCCAT
26611c0g1i2R	TTTGGGGTCTAGGTTTGCTG
25661c1g1i3F	AAGCTCAAGTTGCCAATGCT
25661c1g1i3R	TTGCTTCATCCAACCAATGA
25958c0g1i2F	GACGGAGCTCTTCAGTGGAC
25958c0g1i2R	TACTCCTGTTGGCCGGTTAG
CBTS2aF	ATGAGTCAATCAATTTCTCCATTAATC
CBTS2aR	TCATATGTCGACAGATTCGACA
CBTSProbeF	CATTCCAAAATGTGAGGAGTTAT
CBTSProbeR	TGCTTCTTGCTTTGAAACTGC
eIF4E_C_F	GGATCCACGAAAATGGCAGAGGAAGC
eIF4E_C_R	GAGCTCGCTAATGTCTATAAACTTTCCAGTCCA
cembPro_F	AGCCACAAAAGAGATGAAACC
cembPro_R	CATACGGATCCTCTTTCTCTCGCCAAACGAG

Table A2.2 Annotations of genes within scaffolds expected to be part of VAM

Gene ID	Annotation
Nitab4.5_0007068g0010.1	Bifunctional inhibitor
Nitab4.5_0007068g0020.1	-
Nitab4.5_0007068g0030.1	Acyl-CoA N-acyltransferase
Nitab4.5_0007068g0040.1	Chlorophyll a/b binding protein
Nitab4.5_0007068g0050.1	Histone H2A
Nitab4.5_0007068g0060.1	-

Table A2.2 continued

Nitab4.5_0007068g0070.1	Bifunctional inhibitor
Nitab4.5_0007068g0080.1	-
Nitab4.5_0008902g0010.1	Gelsolin domain
Nitab4.5_0008902g0020.1	-
Nitab4.5_0008902g0030.1	RNA-binding protein
Nitab4.5_0013033g0010.1	Protein kinase
Nitab4.5_0013033g0020.1	Peptidase S9
Nitab4.5_0008971g0010.1	Ribosomal protein L18/L5
Nitab4.5_0008971g0020.1	Protein of unknown function
Nitab4.5_0008971g0030.1	-
Nitab4.5_0008971g0040.1	2Fe-2S ferredoxin-type domain containing protein
Nitab4.5_0008971g0050.1	-
Nitab4.5_0008971g0060.1	Copper amine oxidase
Nitab4.5_0001441g0010.1	HIP116
Nitab4.5_0001441g0020.1	HR-like lesion-inducer
Nitab4.5_0001441g0030.1	-
Nitab4.5_0001441g0040.1	AP2/ERF domain containing protein
Nitab4.5_0001441g0050.1	AP2/ERF domain containing protein
Nitab4.5_0001441g0060.1	AP2/ERF domain containing protein
Nitab4.5_0001441g0070.1	Phosphate permease
Nitab4.5_0001441g0080.1	Myb domain containing protein
Nitab4.5_0001441g0090.1	11-S seed storage protein
Nitab4.5_0001441g0100.1	Putative S-adenosyl-L-methionine-dependent methyltransferase
Nitab4.5_0001441g0110.1	Regulator of chromosome condensation
Nitab4.5_0001441g0120.1	Pleckstrin
Nitab4.5_0001441g0130.1	11-S seed storage protein
Nitab4.5_0001441g0140.1	BED-type predicted
Nitab4.5_0001441g0150.1	-
Nitab4.5_0003843g0010.1	Protein of unknown function
Nitab4.5_0003843g0020.1	Concanavalin A-like lectin/glucanase
Nitab4.5_0003843g0030.1	-
Nitab4.5_0003843g0040.1	Calcium/calmodulin-dependent/calcium-dependent protein kinase
Nitab4.5_0003843g0050.1	Serine/threonine-protein kinase
Nitab4.5_0003843g0060.1	Leucine rich repeat 4
Nitab4.5_0002177g0010.1	Uncharacterized protein

Table A2.2 continued

Nitab4.5_0002177g0020.1	Uncharacterized protein
Nitab4.5_0002177g0030.1	Protein of unknown function
Nitab4.5_0002177g0040.1	-
Nitab4.5_0002177g0050.1	Leucine rich repeat 4 containing protein
Nitab4.5_0002177g0060.1	-
Nitab4.5_0002177g0070.1	Serine/threonine- / dual specificity protein kinase
Nitab4.5_0002177g0080.1	Protein of unknown function
Nitab4.5_0002177g0090.1	-
Nitab4.5_0002177g0100.1	Concanavalin A-like lectin/glucanase
Nitab4.5_0002177g0110.1	Concanavalin A-like lectin/glucanase
Nitab4.5_0002177g0120.1	Calcium/calmodulin-dependent/calcium-dependent protein kinase
Nitab4.5_0002177g0013.1	NAF/FISL domain containing protein
Nitab4.5_0002177g0140.1	Ankyrin repeat-containing protein
Nitab4.5_0002177g0150.1	Transcription factor TGA
Nitab4.5_0002177g0160.1	CCAAT-binding transcription factor
Nitab4.5_0002177g0170.1	Protein of unknown function
Nitab4.5_0003711g0010.1	Transcription factor TGA like domain
Nitab4.5_0003711g0020.1	-
Nitab4.5_0003711g0030.1	Myb domain containing protein
Nitab4.5_0003711g0040.1	-
Nitab4.5_0003711g0050.1	Protein of unknown function
Nitab4.5_0003711g0060.1	-
Nitab4.5_0003711g0070.1	Protein of unknown function
Nitab4.5_0001375g0010.1	PMR5 N-terminal domain containing protein
Nitab4.5_0001375g0020.1	-
Nitab4.5_0001375g0030.1	-
Nitab4.5_0001375g0040.1	Lipase
Nitab4.5_0001375g0050.1	Lipase
Nitab4.5_0001375g0060.1	-
Nitab4.5_0001375g0070.1	-
Nitab4.5_0001375g0080.1	P-loop containing nucleoside triphosphate hydrolase
Nitab4.5_0001375g0090.1	phosphoprotein PP28
Nitab4.5_0001375g0100.1	-
Nitab4.5_0001375g0110.1	-
Nitab4.5_0001375g0120.1	Cell wall/choline-binding repeat
Nitab4.5_0001375g0130.1	Protein of unknown function
Nitab4.5_0001375g0140.1	-

Table A2.2 continued

Nitab4.5_0001375g0150.1	-
Nitab4.5_0001375g0160.1	AP2/ERF domain
Nitab4.5_0001375g0170.1	-
Nitab4.5_0001375g0180.1	-
Nitab4.5_0003487g0010.1	-
Nitab4.5_0003487g0020.1	-
Nitab4.5_0003487g0030.1	Transposase
Nitab4.5_0003487g0040.1	-
Nitab4.5_0003487g0050.1	-
Nitab4.5_0003487g0060.1	Galactose oxidase
Nitab4.5_0003487g0070.1	-
Nitab4.5_0003487g0080.1	Protein of unknown function
Nitab4.5_0003487g0090.1	-
Nitab4.5_0003487g0100.1	Dormancyauxin associated
Nitab4.5_0003487g0110.1	Protein of unknown function
Nitab4.5_0002814g0010.1	Chlorophyll a/b binding protein
Nitab4.5_0002814g0020.1	Histone-lysine N-methyltransferase
Nitab4.5_0002814g0030.1	Chlorophyll a/b binding protein
Nitab4.5_0002814g0040.1	Chlorophyll a/b binding protein
Nitab4.5_0002814g0050.1	Chlorophyll a/b binding protein
Nitab4.5_0002814g0060.1	UDP-N-acetylglucosamine transferase subunit ALG14
Nitab4.5_0002814g0070.1	Cc-nbs-lrr, resistance protein
Nitab4.5_0002814g0080.1	Receptor like kinase, RLK
Nitab4.5_0002814g0090.1	Protein of unknown function
Nitab4.5_0002814g0100.1	Cc-nbs-lrr, resistance protein
Nitab4.5_0002814g0110.1	Phosphatidylinositol 4-kinase IPR015433
Nitab4.5_0002814g0120.1	Eukaryotic translation initiation factor 4E
Nitab4.5_0002814g0130.1	Lipid A export ATP-binding_permease protein
Nitab4.5_0002814g0140.1	Unknown Protein
Nitab4.5_0002814g0150.1	-
Nitab4.5_0002814g0160.1	Outward rectifying potassium channel
Nitab4.5_0002210g0010.1	Protein of unknown function
Nitab4.5_0002210g0020.1	MADS box transcription factor
Nitab4.5_0002210g0030.1	C2 domain-containing protein
Nitab4.5_0002210g0040.1	Mitochondrial import receptor subunit TOM22
Nitab4.5_0002210g0050.1	Protein of unknown function

Table A2.2 continued

Nitab4.5_0002210g0060.1	TPR domain protein
Nitab4.5_0002210g0070.1	Chlorophyll a-b binding protein
Nitab4.5_0002210g0080.1	Protein of unknown function
Nitab4.5_0002210g0090.1	-
Nitab4.5_0002210g0100.1	Signal peptidase complex subunit 1
Nitab4.5_0002210g0110.1	Chlorophyll a-b binding protein
Nitab4.5_0002210g0120.1	Chlorophyll a-b binding protein
Nitab4.5_0002210g0130.1	-
Nitab4.5_0003331g0010.1	Bicoid-interacting 3 domain-containing protein (Fragment)
Nitab4.5_0003331g0020.1	DNA helicase
Nitab4.5_0003331g0030.1	-
Nitab4.5_0003331g0040.1	Root cap protein 3 (Fragment)
Nitab4.5_0003331g0050.1	-
Nitab4.5_0003331g0060.1	Cell number regulator 10
Nitab4.5_0003331g0070.1	Protein of unknown function
Nitab4.5_0003331g0080.1	-
Nitab4.5_0003331g0090.1	Cellulose synthase
Nitab4.5_0003331g0100.1	Protein of unknown function
Nitab4.5_0003331g0110.1	Protein of unknown function
Nitab4.5_0003331g0120.1	Protein of unknown function
Nitab4.5_0003331g0130.1	Phosphatidate cytidylyltransferase
Nitab4.5_0003331g0140.1	Flotillin domain protein
Nitab4.5_0003331g0150.1	Magnesium transporter MRS2-I
Nitab4.5_0003268g0010.1	Ankyrin repeat family protein
Nitab4.5_0003268g0020.1	Protein of unknown function
Nitab4.5_0003268g0030.1	Kinesin like protein
Nitab4.5_0003268g0040.1	Cc-nbs-lrr, resistance protein
Nitab4.5_0003268g0050.1	Cc-nbs-lrr, resistance protein
Nitab4.5_0003268g0060.1	Cc-nbs-lrr, resistance protein
Nitab4.5_0003268g0070.1	Protein of unknown function
Nitab4.5_0003268g0080.1	Receptor like kinase
Nitab4.5_0003268g0090.1	Receptor like kinase
Nitab4.5_0003268g0100.1	Ethylene-responsive transcription factor 4
Nitab4.5_0003268g0110.1	-
Nitab4.5_0003268g0120.1	Cc-nbs-lrr, resistance protein

Table A2.2 continued

Nitab4.5_0010227g0010.1	Cc-nbs-lrr, resistance protein
Nitab4.5_0006402g0010.1	Cc-nbs-lrr, resistance protein
Nitab4.5_0002507g0010.1	Zinc finger family protein
Nitab4.5_0002507g0020.1	Protein of unknown function
Nitab4.5_0002507g0030.1	-
Nitab4.5_0002507g0040.1	-
Nitab4.5_0002507g0050.1	Zinc finger family protein
Nitab4.5_0002507g0060.1	Zinc finger family protein
Nitab4.5_0002507g0070.1	Cyclic phosphodiesterase
Nitab4.5_0002507g0080.1	DUF1264 domain protein
Nitab4.5_0002507g0090.1	Peptide transporter
Nitab4.5_0002507g0100.1	Protein-tyrosine kinase 6
Nitab4.5_0002507g0110.1	Transposase (Fragment)
Nitab4.5_0002507g0120.1	Protein of unknown function
Nitab4.5_0002507g0130.1	GDSL esterase_lipase
Nitab4.5_0002458g0010.1	Zinc finger family protein
Nitab4.5_0002458g0020.1	BES1_BZR1 homolog protein 4
Nitab4.5_0002458g0030.1	Transmembrane water channel aquaporin Z
Nitab4.5_0002458g0040.1	Heat stress transcription factor A3-type
Nitab4.5_0002458g0050.1	DNA binding protein
Nitab4.5_0002458g0060.1	Ubiquitin domain containing 1
Nitab4.5_0002458g0070.1	-
Nitab4.5_0002458g0080.1	Protein of unknown function
Nitab4.5_0002458g0090.1	Receptor like protein
Nitab4.5_0002458g0100.1	Receptor like kinase
Nitab4.5_0002458g0110.1	Zinc finger family protein
Nitab4.5_0002458g0120.1	Ankyrin repeat family protein-like
Nitab4.5_0002458g0130.1	Glycogen synthase kinase
Nitab4.5_0005928g0010.1	Receptor like kinase
Nitab4.5_0005928g0020.1	Accelerated cell death 6 (Fragment)
Nitab4.5_0004518g0010.1	Protein of unknown function
Nitab4.5_0006329g0010.1	Receptor like kinase
Nitab4.5_0001037g0010.1	-
Nitab4.5_0001037g0020.1	Serine_threonine-protein kinase
Nitab4.5_0008217g0010.1	Prenylcysteine oxidase 1

Table A2.3 Viral observations

Transgenic plant,T_0	PVY Sym	PVY ELISA Assay	PVY Immunostrip Assay	Transgenic plant,T_0	PVY Sym.	PVY ELISA Assay	PVY Immunostrip Assay
OX-16	+	2.981	+	TN86(1) VC	-	0.168	-
OX-19	+	1.043	+	TN86(2) VC	-	0.211	-
OX-20	+	0.576	+	TN86(3) VC	-	0.044	-
OX-21	+	0.95	+	TN86(4) VC	-	0.174	-
OX-22	+	1.306	+	TN86(5) VC	-	0.309	-
OX-27	+	0.267	+	TN86(6) VC	-	0.207	-
OX-28	+	0.303	+	TN86(7) VC	-	0.175	-
OX-29	+	0.27	+	TN86(8) VC	-	0.155	-
OX-41	+	0.759	+	TN86(9) VC	-	0.291	-
OX-42	+	0.628	+	TN86(10) VC	-	0.2	-
OX-43	+	0.855	+				
OX-45	+	0.306	+				
OX-48	+	0.414	+				
OX-49	+	0.352	+				

Table showing results of phenotypic observation and immunochemical analysis of 14 T0 35S::eiF4E1.S (OX) and 10 T0 vector control (VC) plants. Positive (+) scoring of PVY symptoms was based on observation of mosaic patterning of leaves.

Table A3.1 Contigs and their annotations

Contig ID	Annotation
114525g1i1	none
121486g1i1	none
123580g10i1	none
125853g1i1	none
125998g2i1	none
126474g1i2	none
1271g2i1	none

Table A3.1 continued

127675g2i4	none
128639g6i2	none
128639g6i7	none
128639g6i8	none
128946g2i2	none
129003g5i2	none
129436g4i1	none
129463g3i12	none
129463g3i2	none
129463g3i6	none
129463g3i9	none
129653g3i1	none
130041g10i1	none
130086g3i2	none
131019g1i3	none
132205g2i1	none
132205g2i2	none
132205g2i4	none
132205g2i5	none
132209g2i2	PREDICTED: putative late blight resistance protein
132209g2i3	PREDICTED: putative late blight resistance protein
133686g10i1	none
133725g3i5	none
133814g8i8	none
134479g1i5	none
134941g1i3	none
134994g1i6	none
134994g1i7	none
136244g2i2	none
136452g7i1	none
136458g1i2	none
136662g7i1	none

137017g2i6	none
137129g6i2	Polyprotein, 3'-partial, putative
137710g2i2	none
138113g4i1	none

Table A3.1 continued

139654g6i1	none
139677g2i4	PREDICTED: putative late blight resistance protein homolog
139856g1i1	uncharacterized protein LOC102581414
139856g1i3	PREDICTED: uncharacterized protein LOC101256790
139856g1i5	PREDICTED: uncharacterized protein LOC101256790
139856g1i6	PREDICTED: uncharacterized protein LOC101256790
139948g2i3	none
140354g1i1	none
140633g3i7	none
140832g1i7	none
141009g1i5	none
142205g5i2	none
142205g5i4	none
142993g3i1	modification methylase, partial
142993g3i2	none
142993g3i3	none
142993g1i1	hypothetical protein CICLEv10003098mg
142993g1i2	hypothetical protein CICLEv10003098mg
143727g3i3	PREDICTED: uncharacterized protein LOC101243773
143775g5i1	PREDICTED: putative NPIP-like protein LOC613037-like
147440g1i3	endonuclease/exonuclease/phosphatase family protein [Medicago truncatula]
165520g1i1	none
171414g1i1	none
171414g1i2	none
55224g1i1	none
55965g1i1	none
60182g2i1	none
60182g2i2	none
116350g1i1	none
122731g2i1	none
123580g16i1	none

124432g5i2	unnamed protein product [Coffea canephora]
124668g1i1	none
126444g2i1	none
126444g2i2	none
128255g6i2	none
128259g3i1	none

Table A3.1 continued

129003g5i1	none
129436g2i1	none
129539g5i2	none
129653g3i2	none
129773g1i1	none
130080g1i1	none
130086g3i7	PREDICTED: putative leucine-rich repeat receptor-like protein kinase At2g19210-like [Solanum tuberosum]
131313g2i2	Zinc knuckle family protein [Solanum bulbocastanum]
132149g5i1	none
133329g3i2	none
133620g8i1	none
137165g4i1	none
138502g4i3	hypothetical protein VITISV006955 [Vitis vinifera]
138863g3i1	none
140362g5i1	PREDICTED: uncharacterized protein LOC101244259 [Solanum lycopersicum]
140491g2i1	PREDICTED: uncharacterized protein LOC101259563 [Solanum lycopersicum]
140633g3i1	none
141009g1i8	none
141702g1i1	none
142205g5i6	none
142630g1i5	none
143727g3i6	PREDICTED: uncharacterized protein LOC101255821 [Solanum lycopersicum]
113682g1i1	none
115048g1i1	none
115677g1i1	none
120922g1i2	PREDICTED: BTB/POZ domain-containing protein At5g03250-like [Solanum lycopersicum]
121166g2i3	none

122174g1i1	none
122174g1i3	none
122515g3i1	none
124359g9i1	none
125921g1i1	PREDICTED: uncharacterized protein LOC103439082, partial [Malus domestica]
126004g1i1	none
126184g2i4	PREDICTED: uncharacterized protein LOC102665746
127675g2i2	none

Table A3.1 continued

128255g6i1	none
128573g1i2	none
128573g1i3	excinuclease ABC subunit B [Bavariicoccus seileri]
128573g6i1	none
128738g1i3	T4.15
128738g1i5	T4.15
129003g5i4	NB-ARC domain containing protein [Solanum demissum]
129003g5i5	NB-ARC domain containing protein [Solanum demissum]
129714g3i1	putative gag protein [Nicotiana tabacum]
130350g5i1	hypothetical protein [Pannonibacter phragmitetus]
131337g3i1	hypothetical protein VITISV018593 [Vitis vinifera]
133620g6i2	none
133686g10i2	none
133686g10i3	NB-ARC domain containing protein [Solanum demissum]
133686g10i4	PREDICTED: putative NPIP-like protein LOC613037-like [Solanum tuberosum]
133725g3i1	none
134199g3i1	RNA-directed DNA polymerase (Reverse transcriptase)
136244g2i3	none
136719g1i4	none
137017g2i3	PREDICTED: nucleobase-ascorbate transporter 7-like
137017g2i8	PREDICTED: nucleobase-ascorbate transporter 4-like isoform X1
138113g4i2	none
138116g1i1	aspartate carbamoyltransferase regulatory subunit
138611g2i4	none
138611g2i7	none
139743g2i1	hypothetical protein MIMGUmgv1a0045102mg, partial

139743g2i3	hypothetical protein
139743g2i4	PREDICTED: uncharacterized protein LOC104085741 [Nicotiana tomentosiformis]
139856g1i4	PREDICTED: uncharacterized protein LOC104110883 [Nicotiana tomentosiformis]
140586g6i1	none
140633g3i4	none
140861g3i2	PREDICTED: nucleobase-ascorbate transporter 4-like [Nicotiana tomentosiformis]
141375g1i7	none
141958g2i1	hypothetical protein SERLADRAFT469399

Table A3.1 continued

143727g3i4	PREDICTED: uncharacterized protein LOC104107118 [Nicotiana tomentosiformis]
143727g3i7	PREDICTED: uncharacterized protein LOC101255821 [Solanum lycopersicum]
145725g1i1	none
145725g1i2	none
177531g1i1	none
85484g1i1	PREDICTED: (E,E)-germacrene B synthase-like [Nicotiana tomentosiformis]
97252g1i1	PREDICTED: dCTP pyrophosphatase 1-like [Nicotiana tomentosiformis]

Table A3.2 Primers used to verify 26 PCR markers

Contig ID	Forward Primer	Reverse Primer
129463g3i2	TGCTTGCTGCATTTTGTTTC	CACGAGGACCGGTAACTGAT
132205g2i1	ACTGTTGGATCCACCGAAAG	TGCAGCACCCTCTTTGACTA
114942g1i1	TCTCAATTATGCGAGCGAAA	AGCCCTAATTGCCTTCCCTA
130041g10i1	CAACCATGGCATCCCTATCT	AGCTTCGGACCAATGAGAGA
136452g7i1	CTCATACCTCCCATGCTGGT	CGCCGTGTAGTTTTCTCGTT
165520g1i1	GGGGAACACTTGGTGTGAAT	CCCCTGTAACTCCGACGTAA
142993g3i3	AACCCAGAATGGACATGAGC	TGTCCCACATTGGAAAAATG
134479g1i5	CATGACTGCTGGTCATTTCG	GTGCGCAACCAAAGCTTATT
137710g2i2	GCAGAAACAATGAGCCAAAA	CGCATCTTCTCATTGCACAC
60182g2i1	CACCCCTGATTCTTCTCGTC	TTGACACTCGCCTGTGTTTC
139743g2i4	ACCTCAGCTGCTTCTTGGAA	TGCTGTCAACAGGAGGAATG
134994g1i7	GAGGAAGTTACGCAGGGTGA	TGGTCAATGCAAACCAAAGA

1271g2i1	CGACAGAAGGCACTTTCACA	TGCTGTCGTTGGAACTTGAG
131019g1i3	GAAAGATGCTCCGCCAATAG	CAGAGCCGTCCCAGAAGTAG
134941g1i3	TTTAGCGCGTTTTTAGGGAAT	TTATCCAATGTGGCTGCAAA
136458g1i2	AAATCACCCATGGGAACAAA	TATATCCGCCCCTGCTTATG
137017g2i6	CGGCAGTGGAGCTAACAAAT	CTGCTTCCAACTCGTGTCAA
136662g7i1	CGTTATCAGCACGATGCTCT	ACATCGATCACGCAACACTG
139654g6i1	CTCGGTGGTGTAGGAGGTGT	TGCTCCAGTAACGGCTCTTT
121166g2i3	GCTTTGGGTTGGAACTGTGT	AAGCAATAACTCGCCGAAAA
126474g1i2	ATGTTGTTGCTGCTTGTTGC	TTGAGTGCAGGTCCAGTGAC
126444g2i2	TTTCAGAACTTGAGGGGAGA	AACCAACCGAAAACCAACAG
140362g5i1	ATTTTTCCAAACTGCGCAAC	GCTTCATTTCCACCCCCTAT
122515g3i1	CGACCCAATACCATTTTCGT	CGAACCCAAGAAAACCTTCA
126184g2i4	TGAGCTCGATTGGTTTTTCC	AGCTTACTGGGCCCAACTTT

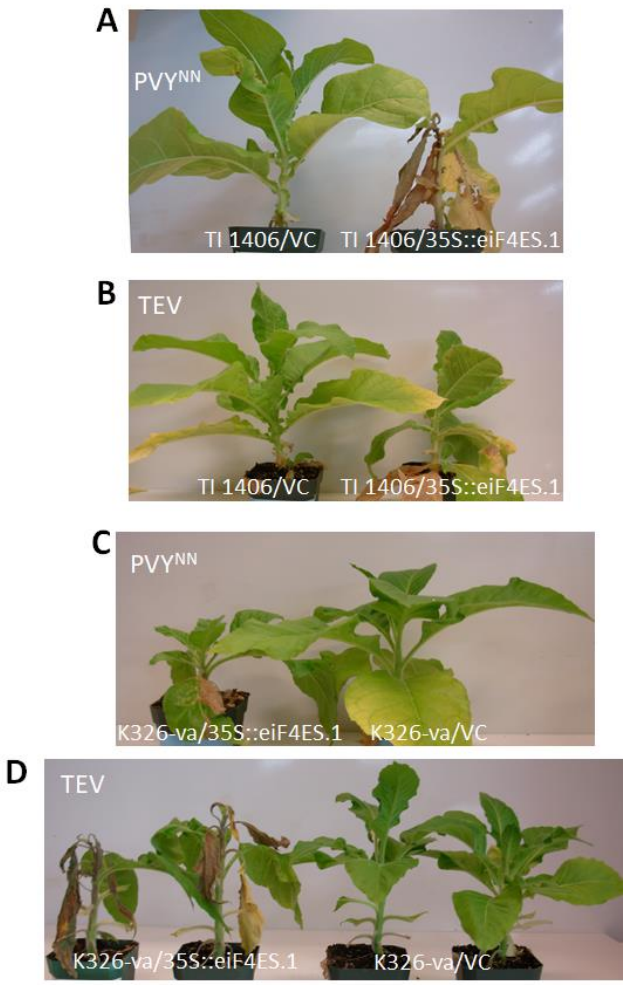


Figure A2.1 Typical examples of potyvirus infection of TI 1406 and K326-*va* plants transformed with 35S::eiF4E1.S construct or vector control (VC). Pictures were taken 14 days post-infection with PVY^{NN} or TEV.