Check for updates

Received: 9 December 2020 | Revised: 8 March 2021 | Accepted: 13 March 2021

DOI: 10.1002/pld3.321

ORIGINAL RESEARCH

A consensus on the Aquaporin Gene Family in the Allotetraploid Plant, *Nicotiana tabacum*

Michael Groszmann[1] | Annamaria De Rosa[1] | Jahed Ahmed[2] | François Chaumont[2] | John R. Evans[1]

[1]ARC Centre of Excellence for Translational Photosynthesis, Research School of Biology, Australian National University, Canberra, ACT, Australia

[2]Louvain Institute of Biomolecular Science and Technology, UCLouvain, Louvain-la-Neuve, Belgium

Correspondence
Michael Groszmann, ARC Centre of Excellence for Translational Photosynthesis, Research School of Biology, Australian National University, Canberra, ACT, Australia.
Email: michael.groszmann@gmail.com; michael.groszmann@anu.edu.au

Funding information
Australian Research Council Center of Excellence for Translational Photosynthesis, Grant/Award Number: CE140100015; Belgian National Fund for Scientific Research (FNRS, FRFC T.0038.18F).

Abstract

Aquaporins (AQPs) are membrane-spanning channel proteins with exciting applications for plant engineering and industrial applications. Translational outcomes will be improved by better understanding the extensive diversity of plant AQPs. However, *AQP* gene families are complex, making exhaustive identification difficult, especially in polyploid species. The allotetraploid species of *Nicotiana tabacum* (Nt; tobacco) plays a significant role in modern biological research and is closely related to several crops of economic interest, making it a valuable platform for AQP research. Recently, De Rosa et al., (2020) and Ahmed et al., (2020), concurrently reported on the AQP gene family in tobacco, establishing family sizes of 76 and 88 members, respectively. The discrepancy highlights the difficulties of characterizing large complex gene families. Here, we identify and resolve the differences between the two studies, clarify gene models, and yield a consolidated collection of 84 members that more accurately represents the complete NtAQP family. Importantly, this consensus NtAQP collection will reduce confusion and ambiguity that would inevitably arise from having two different descriptive studies and sets of NtAQP gene names. This report also serves as a case study, highlighting and discussing variables to be considered and refinements required to ensure comprehensive gene family characterizations, which become valuable resources for examining the evolution and biological functions of genes.

KEYWORDS

Aquaporins, gene family characterization, gene structure and evolution, major intrinsic proteins, orthologs, phylogenetics

1 | INTRODUCTION

1.1 | AQUAPORINS – channel proteins with complex gene families and versatile applications

Aquaporins (AQPs) constitute a major family of membrane-spanning channel proteins that selectively facilitate the passive bidirectional movement of substrates across biological membranes. AQPs are found across all phylogenetic kingdoms but are by far the most diversified in plants, transporting a range of solutes essential for numerous plant processes including, water relations, growth, stress responses, nutrient uptake, and photosynthesis (Chaumont & Tyerman, 2017). Plant AQPs are a complex gene family comprised of several subfamilies, some of which only exist in older plant lineages (e.g., mosses,

Michael Groszmann and Annamaria De Rosa are joint senior authors.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *Plant Direct* published by American Society of Plant Biologists and the Society for Experimental Biology and John Wiley & Sons Ltd.

green algae, and diatoms) and others having emerged in higher order plants (Groszmann et al., 2017; Laloux et al., 2018). Among angiosperms, the AQP family is typically comprised of PIPs, TIPs, NIPs, SIPs, and XIPs, with each subfamily characteristically localizing to different subcellular membranes and transporting somewhat unique sets of substrates. A series of subtypes and a high multiplicity of isoforms have evolved in plants, particularly in the PIPs, TIPs, and NIPs, with AQP families in angiosperms ranging from 30 to 120 members (Groszmann et al., 2017; Laloux et al., 2018). This rich diversification, assorted complement of transported substrates, and an increasing understanding of the many developmental and stress-responsive processes that AQPs are involved in has resulted in a keen interest in targeting AQPs for engineering more resilient and productive plants, as well as for use in biomimetic membranes for industrial filtration applications (Hélix-Nielsen, 2018; Tang et al., 2015). Expanding insight into the diversity and native biological roles of plant AQPs will improve their successful manipulation for translational outcomes.

1.2 | Genomic sequencing and gene family characterizations

From their inception, whole-genome datasets have been used to identify and examine gene families. The rapidly growing collection of fully sequenced plant genomes, including many crop species, represent incredible opportunities for comparative analysis. Provenance, expansion, and evolutionary history of gene families across species can inform us about biological functions.

Numerous studies characterizing *AQP* gene families in plants have been reported. These cover a wide range of species and are contributing valuable insight into the evolution and diversity of plant AQPs. Coupling these with transcriptomic and functional studies is helping to reveal the many native roles of plant AQPs as well as identifying possible targets that could be manipulated for improving crop productivity.

Understanding AQP biology and harnessing the potential of their diversity for engineering efforts relies foremost on the identification of complete gene families, with accurate gene and protein models that allow for the correct assignment of homologs and/or orthologs. Accurate curation of large gene families is time-consuming, laborious, and requires an understanding of the strengths and limitations of datasets and methods being used. Analysis often commences with an iterative process of manual and computational analysis, often combining numerous BLAST searches to identify members of a given family. This is followed by the manual curation of gene models that are typically validated against RNA-seq data. The encoded proteins can also be validated for the presence of conserved signature motifs, secondary/tertiary conformation and, if available, against curated orthologs from closely related species.

Characterizing AQP gene families can be quite complex, given the occurrence of several subfamilies that are further divided into subgroups each with numerous isoforms (Groszmann et al., 2017; Laloux et al., 2018). This is made additionally difficult when examining polyploid species and their multiplicity in genomic content. Recently two concurrently published studies, De Rosa et al. (2020) and Ahmed et al. (2020), characterized the AQP gene family in the allotetraploid species of *Nicotiana tabacum*.

Tobacco is a popular model species capable of scaling from the laboratory to the field. It belongs to the Solanaceae family of angiosperms and is therefore closely related to crops of economic interest such as tomatoes, potatoes, eggplants, and peppers. Tobacco itself has renewed commercial applications in the biofuel and plant-based pharmaceutical sectors. The Nicotiana genus, to which tobacco belongs, has interesting attributes for studying gene family evolution. Its taxonomy is well characterized with approximately 75 species organized into several sections; several key members of which have whole-genome sequencing data available (Bombarely et al., 2012; Schiavinato et al., 2019, 2020; Sierro et al., 2014; Xu et al., 2017). Polyploids formed by interspecific hybridization reside within five of the sections (Clarkson et al., 2004; Leitch et al., 2008). The inception of these polyploids differs in both age and relatedness of the contributing parental species, providing opportunities to examine gene family changes at different stages of polyploid evolution (Leitch et al., 2008). For example, the allotetraploid *Nicotiana benthamiana* (section Suaveolentes), commonly used in labs, emerged ~10 MYA and is in an advanced stage of long-term genome diploidization with extensive rearrangements of its subgenomes (Schiavinato et al., 2020). Whereas, the allotetraploid *N. tabacum*, which arose from a hybridization event between *N. sylvestris* and *N. tomentosiformis*, only emerged ~0.2 MYA and is still in a highly "duplicated" state (Edwards et al., 2017; Sierro et al., 2014). Accordingly, both De Rosa et al. (2020) and Ahmed et al. (2020) identified *N. tabacum* as having a large AQP gene family, currently, second only to the allopolyploid *Brassica napus* (Canola) (Sonah et al., 2017; Yuan et al., 2017). De Rosa et al. established 76 members of the NtAQP family, while Ahmed et al. reported 88 full-length *AQP* genes. The discrepancy highlights the difficulties of characterizing large complex gene families. Here, we compare the two sets of *NtAQP* genes from De Rosa *et al* and Ahmed *et al*, clarify the discrepancies, and arrive at a consensus collection more accurately representing a complete *NtAQP* gene family to better guide evolutionary and functional studies.

2 | RESULTS AND DISCUSSION

2.1 | Determining an assured consensus NtAQP gene family

Each study applied slightly different approaches toward identifying a mostly similar collection of *AQP* genes. Briefly, De Rosa et al. relied on genomic sequence and computed annotations generated via the Solanaceae Genomics Network and performed local BLASTp and BLASTn searches using published potato and tomato AQP genes/protein sequences as queries. Importantly, De Rosa et al. focused on a single cultivar (TN90) for its primary discovery and the K326 cultivar

for secondary confirmation. Gene models were manually refined and validated by comparisons against orthologs and against distribution patterns of mRNA-seq reads mapped to the tobacco TN90 genomic sequence. Ahmed et al. used potato and tomato AQP protein sequences as queries in BLASTp searches through the NCBI portal which draws from the NCBI Reference Sequence (RefSeq) collection of gene/protein annotations (O'Leary et al., 2016). Ahmed et al. relied on RefSeq annotations, which for plant genes are largely predicted by automated computational analysis, and where available, incorporates RNAseq or EST data for validation (O'Leary et al., 2016).

Sequences of all gene/protein models reported in the two studies were imported into Geneious (ver. 11.1.5) and compared using sequence MUSCLE alignments and Neighbor-Joining phylogenetic trees to identify outliers and discrepancies. Flagged annotations were then manually investigated to determine their points of difference. This included scrutinizing the origins/sources of the genetic data, and inspecting splice variants using RNAseq distributions in the RefSeq browser and comparison against models from parental, tomato, and potato genomes. From this analysis, we were able to clarify all discrepancies between the two studies and refine the curation of the *NtAQP* gene family (Table 1; Figure 1).

In De Rosa *et al*., we identified eight significant inconsistencies, all of which were related to unidentified genes (Table 1). Five of these were accounted for due to misses by the genome-wide computational predictions present in the Sol Genome Network (SGN) database, but could be subsequently identified in this study by searching the genomic scaffold sequences directly. Such misses are not inconsistent with the generally lower accuracy of predicted *NtAQP* gene models found with the SGN database (De Rosa et al., 2020). A second copy of *NtPIP2;11* inherited from the *N. sylvestris* parent (*NtPIP2;11bs*), originally thought to be a truncated pseudogene, was found to contain a full-length sequence after further scrutiny of the TN90 genome assembly. *NtTIP2;4t* was not identified as its sequence is not present within the TN90 sequenced genome, and only identified in Ahmed et al. due to their non-selective use of cultivars (see below). *NtTIP1;5t* was not identified as it was missed by the SGN computational predictions and also absent in both parental genomes. Consistent with its reporting in Ahmed et al. we managed to identify *NtTIP1;5t* by BLASTn query of the TN90 scaffold sequences directly. However, further surveys of the parental genomes did not uncover a copy of *NtTIP1;5t*. We did identify a highly homologous copy of *NtTIP1;5t* in the sequenced genome of *Nicotiana otophora* (SGN: scaffold Noto_AWOL01S0468679.1), which belongs to the *Tomentosae* section of the *Nicotiana* genus and thus a close relative to *N. tomentosiformis* (Leitch et al., 2008). Interestingly, early studies into the tobacco lineage implicated *N. otophora* as a part paternal donor to the tobacco genome through an introgressed hybrid with *N. tomentosiformis* (Kenton et al., 1993; Riechers & Timko, 1999). Whole-genome sequencing has since revealed *N. tomentosiformis* as the paternal donor, which largely but not completely dismisses these early studies (Edwards et al., 2017; Sierro et al., 2014). A small fraction of the sequenced tobacco genome is still more similar to *N. otophora* than *N. tomentosiformis*, meaning that the presence of *NtTIP1;5t* in tobacco could be from a genomic introgression of *N. otophora* at some point in the evolution of tobacco. This would be consistent with the phylogenetic isolation of *NtTIP1;5t* with it being the only NtAQP to have no parental copy and an absence of a sister gene (Figure 1) (De Rosa et al., 2020). Although its origins are ambiguous, the *NtTIP1;5t* isoform was assigned the suffix "t" to denote that at the very least its lineage is from the *N. tomentosiformis* containing *Tomentosae* section of the *Nicotiana* genus as opposed to coming from the *N. sylvestris* parent.

The *NtAQP* gene list in Ahmed et al. contained 12 significant inconsistencies (Table 1). Four of these were missed genes not present in the NCBI RefSeq database. One (Ahmed ID: NtPIP1;11) is a pseudogene identified in De Rosa *et al*. The remaining seven are duplicated annotations inflating gene numbers, that are either gene copies from another cultivar, an unsupported splice variant of an already identified gene (Ahmed ID: *NtNIP5;3*) or a duplicated gene containing a likely sequencing mutation (Ahmed ID: *NtPIP1;5*). Duplicate gene copies can be identified through NCBI BLASTp searches, even if using the RefSeq non-redundant database, because multiple copies of the same gene will be found if polymorphisms are present; the most frequent instance being inter-cultivar polymorphisms (e.g., Ahmed ID: NtPIP1;8 TN90 cultivar = NtPIP1;3 Wisconsin 38 cultivar). We identified these cultivar discrepancies by scrutinizing cultivar information embedded within the RefSeq sequence records or searching the linked publication. We also identified 10 minor inconsistencies in the Ahmed et al. *NtAQP* gene list (Table 1). Most of these were variant protein products encoded by unlikely splice variants that were inconsistent with RNA-seq mapping distributions in tobacco, and gene models from the *N. sylvestris* and *N. tomentosiformis* parental genomes and tomato and potato orthologs; web links to browser views graphically representing this information are embedded in the description column of Table 1.

From this comparative analysis, we derived a consensus set of 84 full-length *AQP* genes (plus the six pseudogenes reported in De Rosa et al.), encoding for 86 protein products (two confirmed splice variants each for NtXIP1;7s and 1;7t), that represent a more accurate and complete *NtAQP* gene family (Table 1, Figure 1, Data File S1). The NtAQPs proteins segregate into the five distinct subfamilies common to higher plants, namely the NIPs (20 isoforms), SIPs (5), PIPs (30), TIPs (25), and XIPs (6). The majority of the NtAQPs are present as phylogenetic pairs, which are the sister genes inherited from the parental *N. sylvestris* and *N. tomentosiformis* parental genomes (Figure 1). These are still largely retained due to the limited genome downsizing that has occurred in tobacco given its short evolutionary time frame since inception (0.2 M years) (detailed in De Rosa et al., 2020). Nomenclature was allocated as set out in De Rosa et al. which adheres to the convention that gene names should reflect orthology across species (e.g., *NtPIP1;1* is the ortholog of tomato *SlPIP1;1*). Unfortunately, the haphazard naming of genes across *AQP* gene family studies is prevalent and causing confusion through missed or incorrectly assigned orthology. Our nomenclature aligns the NtAQP naming convention with that of tomato and additionally incorporates suffix identifiers "t" or "s" to denote the parental subgenome origins of the sister genes in tobacco (Figure 1); a convention

TABLE 1 Comparing *NtAQP* gene annotations from De Rosa et al. (2020) and Ahmed et al. (2020) and deriving a consensus list of 84 members that more accurately represents the complete NtAQP family

This study	De Rosa et al. (2020)				Ahmed et al. (2020)			
Consensus NtAQP family	Gene ID[a]	NCBI Accession (1)	NCBI Accession (2)	Status	Gene ID[b]	NCBI Accession (3)	Status	Description of discrepancy
NtPIP1;1s	*NtPIP1;1s*	**BK011392**			*NtPIP1;7*	**NP_001312824.1**		
NtPIP1;1t	*NtPIP1;1t*	**BK011393**			*NtPIP1;8*	**NP_001312222.1**		
	—	—	—	—	*NtPIP1;3*	AAB04757.1		Duplicated annotation. Additional copy NtPIP1;1t from Wisconsin 38 cultivar. Polymorphisms exist compared to cv. TN90.
NtPIP1;2s	*NtPIP1;2s*	**BK011394**			*NtPIP1;12*	**NP_001312721.1**		
NtPIP1;2t	*NtPIP1;2t*	**BK011395**			*NtPIP1;13*	**XP_016510215.1**		
NtPIP1;3s	*NtPIP1;3s*	**BK011396**			*NtPIP1;1*	NP_001313131.1		NP_001313131.1 sequence is from cv. Petit Havana SR1; polymorphisms exist with cv. TN90.
NtPIP1;3t	*NtPIP1;3t*	**BK011397**			*NtPIP1;6*	**XP_016476491.1**		
NtPIP1;5s	*NtPIP1;5s*	**BK011398**			*NtPIP1;4*	**NP_001312189.1**		
	—	—	—	—	*NtPIP1;5*	CAA04750.1		Duplicated annotation. Identical to NtPIP1;5s (DeRosa, 2020) and NtPIP1;4 (Ahmed, 2020), a lab based submission with a likley non-synonomous sequencing error of a strongly conserved functional residue in the widely studied NtAQP1; detailed in De Rosa et al. (2020).
NtPIP1;5t	*NtPIP1;5t*	**BK011399**			*NtPIP1;2*	**XP_016508253.1**		
NtPIP1;7t	*NtPIP1;7t*	**BK011401**			*NtPIP1;9*	**NP_001312921.1**		
NtPIP1;8s	*NtPIP1;8s*	**BK011400**			*NtPIP1;10*	**XP_016458231.1**		
NtPIP2;1s	*NtPIP2;1s*	**BK011402**			*NtPIP2;15*	**NP_001312333.1**		
NtPIP2;1x	*NtPIP2;1x*	**BK011403**			*NtPIP2;13*	NP_001312334.1		Unlikely extended N-terminus. Not supported by orthologs in other Solanaceae species and upstream AUG is an unfavorable Kozak's context.
NtPIP2;2t	*NtPIP2;2t*	**BK011404**			*NtPIP2;16*	**XP_016513533.1**		
NtPIP2;3t	*NtPIP2;3t*	**BK011405**			*NtPIP2;14*	XP_016486700.1		Unlikely extended N-terminus. Not supported by orthologs in other Solanaceae species and upstream AUG is an unfavorable Kozak's context.
NtPIP2;4s	*NtPIP2;4s*	**BK011406**			*NtPIP2;20*	**NP_001311719.1**		
NtPIP2;4t	*NtPIP2;4t*	**BK011407**			*NtPIP2;21*	**NP_001311765.1**		
NtPIP2;5s	*NtPIP2;5s*	**BK011408**			*NtPIP2;12*	**NP_001312276.1**		
NtPIP2;5t	*NtPIP2;5t*	**BK011409**			*NtPIP2;11*	**NP_001311701.1**		
NtPIP2;6s	*NtPIP2;6s*	**BK011410**			*NtPIP2;18*	**NP_001313066.1**		
NtPIP2;6t	*NtPIP2;6t*	**BK011411**			*NtPIP2;17*	**NP_001312464.1**		
NtPIP2;7t	*NtPIP2;7t*	**BK011412**			*NtPIP2;19*	**NP_001313208.1**		

(Continues)

TABLE 1 *Continued*

This study	De Rosa et al. (2020)				Ahmed et al. (2020)			
Consensus NtAQP family	Gene ID[a]	NCBI Accession (1)	NCBI Accession (2)	Status	Gene ID[b]	NCBI Accession (3)	Status	Description of discrepancy
NtPIP2;8s	*NtPIP2;8s*	**BK011413**			*NtPIP2;3*	**NP_001312414.1**		
NtPIP2;8t	*NtPIP2;8t*	**BK011414**			*NtPIP2;2*	**NP_001313091.1**		
	–	–	–	–	*NtPIP2;1*	AAL33586.1		Duplicated annotation. Additional copy of NtPIP2;8t from Petit Havana SR1 cultivar.
NtPIP2;9s	*NtPIP2;9s*	**BK011415**			*NtPIP2;5*	**NP_001312874.1**		
NtPIP2;9t	*NtPIP2;9t*	**BK011416**			*NtPIP2;4*	**NP_001312350.1**		
NtPIP2;11s	*NtPIP2;11s*	**BK011417**			*NtPIP2;6*	**XP_016477641.1**		
NtPIP2;11bs	*NtPIP2;11s*$_{pseudo}$				*NtPIP2;7*	**NP_001313061.1**		Full length gene not identified in De Rosa et al. (2020)—no predicted gene model in Sol Genome dataset.
NtPIP2;11t	*NtPIP2;11t*	**BK011418**			*NtPIP2;8*	**XP_016476355.1**		
NtPIP2;13s	*NtPIP2;13s*	**BK011419**			*NtPIP2;10*	**XP_016494749.1**		
NtPIP2;13t	*NtPIP2;13t*	**BK011420**			*NtPIP2;9*	**NP_001312511.1**		
					NtPIP1;11	XP_016515710.1		Identified as a pseudo gene (NtPIP1;2bs$_{pseudo}$) with C-terminal truncations in De Rosa et al. (2020).
NtTIP1;1s	*NtTIP1;1s*	**BK011426**			*NtTIP1;3*	**NP_001312871.1**		
NtTIP1;1t	*NtTIP1;1t*	**BK011427**			*NtTIP1;2*	NP_001312131.1		NP_001312131.1 sequence is from cv. Bright Yellow 2; polymorphisms exist with cv. TN90.
	–	–	–	–	*NtTIP1;1*	BAF95576.1		Duplicated annotation. Additional copy of NtTIP1;1t from Bright Yellow 2 cultivar.
NtTIP1;2s	*NtTIP1;2s*	**BK011428**			*NtTIP1;6*	**XP_016487055.1**		
NtTIP1;2t	*NtTIP1;2t*	**BK011429**			*NtTIP1;5*	**XP_016501711.1**		
NtTIP1;3s	*NtTIP1;3s*	**BK011431**			*NtTIP1;9*	**XP_016450483.1**		
NtTIP1;3t	*NtTIP1;3t*	**BK011430**			*NtTIP1;8*	**XP_016495978.1**		
NtTIP1;4t	*NtTIP1;4t*	**BK011432**			*NtTIP1;4*	**XP_016513281.1**		
NtTIP1;5t	–	–	–		*NtTIP1;7*	**XP_016471957.1**		Not identified in De Rosa et al. (2020)—no predicted gene model in Sol Genome dataset.
NtTIP2;1s	*NtTIP2;1s*	**BK011433**	**XP_016462920.1**		–	–		Not identified in Ahmed et al. (2020).
NtTIP2;1t	*NtTIP2;1t*	**BK011434**			*NtTIP2;3*	**XP_016503582.1**		
NtTIP2;2s	*NtTIP2;2s*	**BK011435**			*NtTIP2;7*	**XP_016481958.1**		
NtTIP2;2t	–	–	–		*NtTIP2;6*	**XP_016445220.1**		Not identified in De Rosa et al. (2020)—no predicted gene model in Sol Genome dataset.
NtTIP2;3s	*NtTIP2;3s*	**BK011436**			*NtTIP2;9*	**XP_016481922.1**		

(Continues)

TABLE 1 *Continued*

This study	De Rosa et al. (2020)				Ahmed et al. (2020)			
Consensus NtAQP family	Gene ID[a]	NCBI Accession (1)	NCBI Accession (2)	Status	Gene ID[b]	NCBI Accession (3)	Status	Description of discrepancy
NtTIP2;3t	*NtTIP2;3t*	**BK011437**			*NtTIP2;8*	**NP_001312940.1**		
	–	–	–	–	*NtTIP2;10*	P24422.2		Duplicated annotation. Additional copy of NtTIP2;3s from Wisconsin 38 cultivar.
NtTIP2;4s	*NtTIP2;4s*	**BK011438**			*NtTIP2;5*	**XP_016515893.1**		
NtTIP2;4t	–	–	–		*NtTIP2;4*	**XP_016480756.1**		Not identified in De Rosa et al. (2020); not present in the TN90 cultivar genomic sequence. Identified in Ahmed et al. (2020) using MSK326 cultivar.
NtTIP2;5s	*NtTIP2;5s*	**BK011439**			*NtTIP2;1*	**NP_001312646.1**		
NtTIP2;5t	*NtTIP2;5t*	**BK011440**			*NtTIP2;2*	**XP_016495734.1**		
NtTIP3;1s	*NtTIP3;1s*	**BK011441**			*NtTIP3;3*	**XP_016436583.1**		
NtTIP3;1t	*NtTIP3;1t*	**BK011442**			*NtTIP3;4*	**XP_016500896.1**		
NtTIP3;2t	*NtTIP3;2t*	**BK011443**			*NtTIP3;2*	**XP_016491898.1**		
	–	–	–	–	*NtTIP3;1*	XP_016491554.1		Duplicated annotation. Additional copy of NtTIP3;2t from MSK326 cultivar.
NtTIP4;1s	*NtTIP4;1s*	**BK011444**			*NtTIP4;2*	**XP_016441470.1**		
NtTIP4;1t	*NtTIP4;1t*	**BK011445**			*NtTIP4;1*	**NP_001311953.1**		
NtTIP5;1s	*NtTIP5;1s*	**BK011446**			*NtTIP5;2*	**XP_016485861.1**		
NtTIP5;1t	*NtTIP5;1t*	**BK011447**			*NtTIP5;1*	**XP_016462485.1**		
NtNIP1;1s	*NtNIP1;1s*	**BK011376**			–	–		Not identified in Ahmed et al. (2020).
NtNIP1;2s	*NtNIP1;2s*	**BK011377**			*NtNIP1;1*	**XP_016487110.1**		
NtNIP1;2t	*NtNIP1;2t*	**BK011378**	**XP_016445610.1**		*NtNIP1;2*	XP_016445609.1		Extended N-terminal splice variant of NtNIP1;2t; not supported by RNA-seq and models of parental and Solanaceae orthologs. Browser link: https://go.usa.gov/x7cQk
NtNIP2;1s	*NtNIP2;1s*	**BK011379**			*NtNIP2;1*	**XP_016451246.1**		
NtNIP3;1s	*NtNIP3;1s*	**BK011380**			*NtNIP3;2*	**XP_016515586.1**		
NtNIP3;1t	–	–	–		*NtNIP3;1*	**XP_016460638.1**		Not identified in De Rosa et al. (2020)—no predicted gene model in Sol Genome dataset.
NtNIP4;1s	*NtNIP4;1s*	**BK011381**			*NtNIP4;3*	**XP_016491262.1**		
NtNIP4;1t	*NtNIP4;1t*	**BK011382**			*NtNIP4;4*	**XP_016453373.1**		
NtNIP4;2s	*NtNIP4;2s*	**BK011383**			*NtNIP4;6*	**XP_016500017.1**		
NtNIP4;2t	*NtNIP4;2t*	**BK011384**			*NtNIP4;5*	**XP_016456203.1**		
NtNIP4;3s	*NtNIP4;3s*	**BK011385**			*NtNIP4;1*	**XP_016486634.1**		
NtNIP4;3t	–	–	–		*NtNIP4;2*	**XP_016455585.1**		Not identified in De Rosa et al. (2020)—no predicted gene model in Sol Genome dataset.

(Continues)

TABLE 1 *Continued*

This study	De Rosa et al. (2020)				Ahmed et al. (2020)			
Consensus NtAQP family	Gene ID[a]	NCBI Accession (1)	NCBI Accession (2)	Status	Gene ID[b]	NCBI Accession (3)	Status	Description of discrepancy
NtNIP5;1s	*NtNIP5;1s*	**BK011386**			*NtNIP5;2*	**NP_001312819.1**		
NtNIP5;1t	*NtNIP5;1t*	**BK011387**			*NtNIP5;1*	**XP_016470302.1**		
	–	–	–	–	*NtNIP5;3*	XP_016493176.1		Duplicated annotation. Misidentified gene copy of an unlikely splice variant of NtNIP5;1s not supported by RNA-seq distribution patterns. Browser link: https://go.usa.gov/x7cnz
NtNIP6;1s	*NtNIP6;1s*	**BK011388**	**XP_016435921.1**		*NtNIP6;1*	XP_016435920.1		Extended N-terminal splice variant of NtNIP6;1s; not supported by RNA-seq and models of parental and Solanaceae orthologs. Browser link: https://go.usa.gov/x7cn5
NtNIP6;1t	*NtNIP6;1t*	**BK011389**	**XP_016438238.1**		*NtNIP6;2*	XP_016438237.1		Extended N-terminal splice variant of NtNIP6;1t; not supported by RNA-seq and models of parental and Solanaceae orthologs. Browser link: https://go.usa.gov/x7cnR
NtNIP7;1s	*NtNIP7;1s*	**BK011390**			*NtNIP7;2*	**XP_016496646.1**		
NtNIP7;1t	*NtNIP7;1t*	**BK011391**			*NtNIP7;1*	**XP_016509644.1**		
NtNIP8;1s	–	–	–		*NtNIP8;1*	**XP_016468207.1**		Not identified in De Rosa et al. (2020)—no predicted gene model in Sol Genome dataset.
NtNIP8;1t	–	–	**see Additional File 1**		*NtNIP8;2*	XP_016451938.1		Not identified in De Rosa et al. (2020)—no predicted gene model in Sol Genome dataset; see Additional File 1 for sequence of revised model.
NtSIP1;1t	*NtSIP1;1t*	**BK011421**	**XP_016439605.1**		*NtSIP1;1*	XP_016439604.1		C-terminal splice variant of NtSIP1;1t; not supported by RNA-seq and models of parental and Solanaceae orthologs. Browser link: https://go.usa.gov/x7r4m
NtSIP1;2s	*NtSIP1;2s*	**BK011422**	**XP_016461483.1**		–	–		Not identified in Ahmed et al. (2020).
NtSIP1;2t	*NtSIP1;2t*	**BK011423**			*NtSIP1;2*	**XP_016492107.1**		
NtSIP2;1s	*NtSIP2;1s*	**BK011424**	**XP_016510678.1**		–	–		Not identified in Ahmed et al. (2020).
NtSIP2;1t	*NtSIP2;1t*	**BK011425**			*NtSIP2;1*	**XP_016496337.1**		
NtXIP1;6s	*NtXIP1;6s*	**BK011448**			*NtXIP2;1*	**XP_016489264.1**		
NtXIP1;6t	*NtXIP1;6t*	**BK011449**			*NtXIP2;2*	**XP_016488683**		
NtXIP1;7s	*NtXIP1;7s*	**BK011450**			*NtXIP1;2*	**XP_016446694**		β splice variant version—NCBI accession XP_016446695.1.
NtXIP1;7t	*NtXIP1;7t*	**BK011451**			*NtXIP1;1*	**NP_001312796**		β splice variant version—NCBI accession HM475295.1.

, Accurate annotation; , Minor inconsistency with annotation; , Significant inconsistency with annotation.

The consensus gene names and their correct corresponding NCBI accession identification codes are in bold.

NCBI Accession (1): Third-party annotation submissions to NCBI representing curated gene/protein models in De Rosa et al.; (2): NCBI RefSeq records supporting De Rosa et al., TPA submissions in instances of incorrect/unlikely models proposed in Ahmed et al. (2020); (3): NCBI RefSeq records reported in Ahmed et al. (2020). Protein and coding sequences for all 84 AQP members are in Data File S1.

Abbreviation: cv, cultivar.

[a]*NtAQP* gene identifiers as reported in De Rosa et al. (2020).

[b]*NtAQP* gene identifiers as reported in Ahmed et al. (2020).

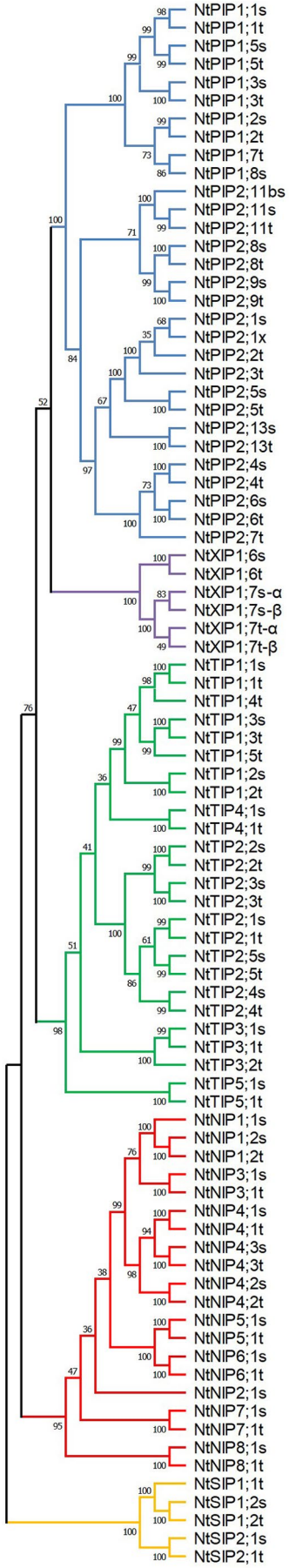


FIGURE 1 Phylogeny of the 86 protein products produced by the consensus 84 genes of the tobacco aquaporin family. Branches are color coded in reference to the five sub-families of PIPs (blue), XIPs (purple), TIPs (green), NIPs (red), and SIPs (Orange). Phylogenetic tree was generated using the neighbor-joining method with pair-wise deletions (via MEGA10) from MUSCLE aligned protein sequences. Confidence levels (%) of branch points generated through bootstrapping analysis (n = 1,000). Suffix identifiers "s" or "t" to denote the sub-genome origins of the sister genes (i.e., "s" = *N. sylvestris* and "t" = *N. tomentosiformis*) that reside as discrete pairings across the phylogeny. The six pseudogenes identified in De Rosa et al. (2020) have not been included in this phylogeny. Protein and coding sequences for all NtAQP members are in Data File S1

customary when naming genes in polyploid species, as for example, in canola and wheat (Adamski et al., 2018; Østergaard & King, 2008).

3 | CONCLUSION

Here, we performed a comparative analysis of two independently collated *NtAQP* gene sets to produce an updated consensus tobacco AQP family. Each study used practises common to many gene family characterisations, and as would be expected, yielded a mostly similar collection of *NtAQP* genes. However, given the slightly different datasets and approaches used, incorrect identifications and missing family members needed to be resolved. The discrepancies reflect the difficulty of characterizing complex gene families, such as the AQPs, especially in polyploid species. Interrogation and consolidation of the two gene sets yielded a more accurate and complete consensus NtAQP family. The comparison also serves to highlight variables to be considered and the refinements required to ensure comprehensive gene family characterizations. These include the use of multiple databases of predicted annotations (if available) in initial BLAST searches and an understanding of the possible outputs, screening of genomic sequences for unannotated genes, careful manual curation of gene and protein models using RNAseq data and established structures of closely related orthologs for validation, and finally naming genes to reflect their orthology.

Importantly, having this consensus *NtAQP* collection will reduce confusion and ambiguity that would inevitably arise from having two different descriptive studies and sets of *NtAQP* gene names. Our refined update to the *NtAQP* gene family can be confidently used for further assessing the evolution of AQPs, and together with the original studies (Ahmed et al., 2020; De Rosa et al., 2020), represent an excellent resource to guide further functional analysis and help transfer basic research to applied outcomes.

4 | METHODS

Sequences of all gene/protein models reported in the two studies were retrieved from NCBI using the NCBI Batch Entrez portal.

Sequences were then imported into Geneious (ver. 11.1.5) and compared using MUSCLE sequence alignments and Neighbor-Joining phylogenetic trees to identify outliers and discrepancies between the two studies. Flagged annotations were then manually investigated to determine points of difference. This included scrutinizing; (a) sequence alignments; (b) origins/sources of the genetic data (e.g., cultivar, predicted gene model, secondary confirmation data); (c) inspecting splice variants and intron/exon gene models using RNAseq distributions in the RefSeq browser (e.g. links to browser shots can be found in Table 1); and (d) comparison of NtAQP models against models from parental, tomato, and potato genomes previously identified/collated in De Rosa et al. (2020). Additional required blast searches were against *N. tabacum* (TN90—TC586—PI 543792; K326—TC 319—PI 552505), *N. tomentosiformis* (Goodsp—TW142—PI 555572), and *N. sylvestris* (Speg—TW136—PI 555569); https://www.ars-grin.gov/npgs/. Final phylogeny for Figure 1 was produced in MEGA 10 (Kumar et al., 2018).

ACKNOWLEDGMENTS

The work was supported by the Australian Research Council Center of Excellence for Translational Photosynthesis (CE140100015) and the Belgian National Fund for Scientific Research (FNRS, FRFC T.0038.18F). J.A. was a research fellow at the Fonds de Formation à la Recherche dans l'Industrie et l'Agriculture.

CONFLICTS OF INTEREST

These authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

M.G and A.D performed the analysis and prepared published data. M.G wrote the manuscript. J.R.E and F.C critically reviewed the manuscript. J.R.E, A.D, F.C, and J.A contributed to discussions on the analysis.

DATA AVAILABILITY STATEMENT

All relevant data can be found within the manuscript and its supporting materials. All sequences are present in the NCBI databases, with their corresponding accession number listed in Table 1. Materials are also available from the corresponding authors upon request.

ORCID

Michael Groszmann https://orcid.org/0000-0002-5015-6156
Annamaria De Rosa https://orcid.org/0000-0002-2610-9149
Jahed Ahmed https://orcid.org/0000-0002-7102-7206
François Chaumont https://orcid.org/0000-0003-0155-7778
John R. Evans https://orcid.org/0000-0003-1379-3532

REFERENCES

Adamski, N. M., Borrill, P., Brinton, J., Harrington, S. A., Marchal, C., Bentley, A. R., Bovill, W. D., Cattivelli, L., Cockram, J., Contreras-Moreira, B., Ford, B., Ghosh, S., Harwood, W., Hassani-Pak, K., Hayta, S., Hickey, L. T., Kanyuka, K., King, J., Maccaferri, M., … Uauy, C. (2020). A roadmap for gene functional characterisation in crops with large genomes: Lessons from polyploid wheat. *eLife*, *9*. http://dx.doi.org/10.7554/elife.55646

Ahmed, J., Mercx, S., Boutry, M., & Chaumont, F. (2020). Evolutionary and predictive functional insights into the aquaporin gene family in the Allotetraploid Plant Nicotiana tabacum. *International Journal of Molecular Sciences*, *21*, 4743.–https://doi.org/10.3390/ijms21134743

Bombarely, A., Rosli, H. G., Vrebalov, J., Moffett, P., Mueller, L. A., & Martin, G. B. (2012). A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research. *Molecular Plant-Microbe Interactions*, *25*, 1523–1530.

Chaumont, F., & Tyerman, S. (2017). *Plant Aquaporins From Transport to Signaling*. Springer International Publishing. https://doi.org/10.1007/978-3-319-49395-4

Clarkson, J. J., Knapp, S., Garcia, V. F., Olmstead, R. G., Leitch, A. R., & Chase, M. W. (2004). Phylogenetic relationships in Nicotiana (Solanaceae) inferred from multiple plastid DNA regions. *Molecular Phylogenetics and Evolution*, *33*, 75–90. https://doi.org/10.1016/j.ympev.2004.05.002

De Rosa, A., Watson-Lazowski, A., Evans, J. R., & Groszmann, M. (2020). Genome-wide identification and characterisation of Aquaporins in Nicotiana tabacum and their relationships with other Solanaceae species. *BMC Plant Biology*, *20*, 1–29.

Edwards, K. D., Fernandez-Pozo, N., Drake-Stowe, K., Humphry, M., Evans, A. D., Bombarely, A., Allen, F., Hurst, R., White, B., Kernodle, S. P., Bromley, J. R., Sanchez-Tamburrino, J. P., Lewis, R. S., & Mueller, L. A. (2017). A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics*, *18*, 448. https://doi.org/10.1186/s12864-017-3791-6

Groszmann, M., Osborn, H. L., & Evans, J. R. (2017). Carbon dioxide and water transport through plant aquaporins. *Plant, Cell and Environment*, *40*, 938–961. https://doi.org/10.1111/pce.12844

Hélix-Nielsen, C. (2018). Biomimetic membranes as a technology platform: Challenges and opportunities. *Membranes*, *8*, 44. https://doi.org/10.3390/membranes8030044

Kenton, A., Parokonny, A. S., Gleba, Y. Y., & Bennett, M. D. (1993). Characterization of the Nicotiana tabacum L. genome by molecular cytogenetics. *Molecular and General Genetics MGG*, *240*, 159–169. https://doi.org/10.1007/BF00277053

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*, 1547–1549. https://doi.org/10.1093/molbev/msy096

Laloux, T., Junqueira, B., Maistriaux, L. C., Ahmed, J., Jurkiewicz, A., & Chaumont, F. (2018). Plant and mammal aquaporins: Same but different. *International Journal of Molecular Sciences*, *19*, 521. https://doi.org/10.3390/ijms19020521

Leitch, I., Hanson, L., Lim, K., Kovarik, A., Chase, M., Clarkson, J., & Leitch, A. (2008). The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). *Annals of Botany*, *101*, 805–814. https://doi.org/10.1093/aob/mcm326

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189

Østergaard, L., & King, G. J. (2008). Standardized gene nomenclature for the Brassica genus. *Plant Methods*, *4*, 1–4. https://doi.org/10.1186/1746-4811-4-10

Riechers, D. E., & Timko, M. P. (1999). Structure and expression of the gene family encoding putrescine N-methyltransferase in Nicotiana

tabacum: New clues to the evolutionary origin of cultivated tobacco. *Plant Molecular Biology*, *41*, 387–401.

Schiavinato, M., Marcet-Houben, M., Dohm, J. C., Gabaldón, T., & Himmelbauer, H. (2020). Parental origin of the allotetraploid tobacco Nicotiana benthamiana. *The Plant Journal*, *102*, 541–554.

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

Sierro, N., Battey, J. N., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S., Peitsch, M. C., & Ivanov, N. V. (2014). The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications*, *5*, 1–9. https://doi.org/10.1038/ncomms4833

Sonah, H., Deshmukh, R. K., Labbé, C., & Bélanger, R. R. (2017). Analysis of aquaporins in Brassicaceae species reveals high-level of conservation and dynamic role against biotic and abiotic stress in canola. *Scientific Reports*, *7*, 1–17. https://doi.org/10.1038/s41598-017-02877-9

Tang, C., Wang, Z., Petrinić, I., Fane, A. G., & Hélix-Nielsen, C. (2015). Biomimetic aquaporin membranes coming of age. *Desalination*, *368*, 89–105. https://doi.org/10.1016/j.desal.2015.04.026

Xu, S., Brockmöller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z., Zhou, W., Kreitzer, C., Stanke, M., Tang, H., Lyons, E., Pandey, P., Pandey, S. P., Timmermann, B., Gaquerel, E., & Baldwin, I. T. (2017). Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proceedings of the National Academy of Sciences*, *114*, 6133–6138. https://doi.org/10.1073/pnas.1700073114

Yuan, D., Li, W., Hua, Y., King, G. J., Xu, F., & Shi, L. (2017). Genome-wide identification and characterization of the aquaporin gene family and transcriptional responses to boron deficiency in Brassica napus. *Frontiers in Plant Science*, *8*, 1336. https://doi.org/10.3389/fpls.2017.01336

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

How to cite this article: Groszmann M, De Rosa A, Ahmed J, Chaumont F, Evans JR. A consensus on the Aquaporin Gene Family in the Allotetraploid Plant, *Nicotiana tabacum*. *Plant Direct*. 2021;5:e00321. https://doi.org/10.1002/pld3.321