frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 August 2016
doi: 10.3389/fpls.2016.01185

# *In silico* Identification and Taxonomic Distribution of Plant Class C GH9 Endoglucanases

*Siddhartha Kundu*[1,2,3]* *and Rita Sharma*[3]*

[1] Department of Biochemistry, Dr. Baba Saheb Ambedkar Medical College & Hospital, New Delhi, India, [2] Mathematical and Computational Biology, Information Technology Research Academy, Media Lab Asia, New Delhi, India, [3] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

The glycoside hydrolase 9 superfamily, mainly comprising the endoglucanases, is represented in all three domains of life. The current division of GH9 enzymes, into three subclasses, namely A, B, and C, is centered on parameters derived from sequence information alone. However, this classification is ambiguous, and is limited by the paralogous ancestry of classes B and C endoglucanases, and paucity of biochemical and structural data. Here, we extend this classification schema to putative GH9 endoglucanases present in green plants, with an emphasis on identifying novel members of the class C subset. These enzymes cleave the $\beta(1 \rightarrow 4)$ linkage between non-terminal adjacent D-glucopyranose residues, in both, amorphous and crystalline regions of cellulose. We utilized non redundant plant GH9 enzymes with characterized molecular data, as the training set to construct Hidden Markov Models (HMMs) and train an Artificial Neural Network (ANN). The parameters that were used for predicting dominant enzyme function, were derived from this training set, and subsequently refined on 147 sequences with available expression data. Our knowledge-based approach, can ascribe differential endoglucanase activity (A, B, or C) to a query sequence with high confidence, and was used to construct a local repository of class C GH9 endoglucanases (*GH9C* = 241) from 32 sequenced green plants.

**Keywords: artificial neural network, carbohydrate binding module, cellulose, endoglucanase, glycoside hydrolase, hidden markov models**

OPEN ACCESS

***Edited by:***
*Manisha Goel,*
*University of Delhi, India*

***Reviewed by:***
*Michael Poidinger,*
*Singapore Immunology Network, Singapore*
*Arthur Gruber,*
*University of São Paulo, Brazil*

***Correspondence:**
*Siddhartha Kundu*
*siddhartha_kundu@yahoo.co.in*
*Rita Sharma*
*rita.genomics@gmail.com*

***Specialty section:***
*This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Plant Science*

***Received:*** *04 April 2016*
***Accepted:*** *22 July 2016*
***Published:*** *12 August 2016*

***Citation:***
*Kundu S and Sharma R (2016) In silico Identification and Taxonomic Distribution of Plant Class C GH9 Endoglucanases. Front. Plant Sci. 7:1185. doi: 10.3389/fpls.2016.01185*

## INTRODUCTION

Cellulose, a straight chain organic polymer of several hundreds of repeating disaccharide units of D-glucopyranose in a $\beta(1 \rightarrow 4)$ glycosidic linkage, is present in the primary cell wall of plants, algae, and oomycetes, and is also a critical component of bacterial biofilms (Updegraff, 1969; Yoshida Y. et al., 2006; Reardon-Robinson et al., 2014; Augimeri et al., 2015). Unlike the $\alpha(1 \rightarrow 4)$ linked glucans of starch (coiled) and glycogen (branched), the $\beta(1 \rightarrow 4)$ bond of cellulose imposes several constraints on its structural conformation, rendering it, inflexible and stiff. Whilst, the bacterial forms are relatively uniform in constitution, plant cell walls are heterogenous, with a mixture of cellulose, hemicelluloses, and lignin (Klemm et al., 2005). The

**Abbreviations:** ANN, Artificial Neural Network; CBM, Carbohydrate Binding Module; GH, Glycoside Hydrolase; HMM, Hidden Markov Model; UID, Uniprot Identifier.

cohesive structure of cellulose is aided, additionally, by a rich network of non-covalent hydrogen bonds between the hydroxyl ($-OH^-$) groups of the glucose moieties of its constituent microfibrils. The resultant macromolecule is stable, and can only be fragmented at elevated temperatures ($>$ 350°C) and pressure, in association with concentrated acids (Agarwal et al., 2012; Paulsen et al., 2014). The presence of cellulose in primary cell walls, whilst protective and strength conferring, is also important in the development and maintenance of bacterial biofilms for host interaction (Rhizobiaceae, Enterobacteriaceae, Acetobacteriaceae, etc.; Augimeri et al., 2015). This stability of cellulose, mandates a breakdown into constituent mono- and oligo-saccharides, prior to major patho-physiological events in plants. In fact, plant development, along with stress adaptor mechanisms, are critically dependent on the digestion of cellulose (del Campillo et al., 2012; Kundu, 2015a).

Endoglucanases, or cellulose hydrolases (EC3.2.1.4), cleave the $\beta\,(1 \rightarrow 4)$ glycosidic linkage of adjacent D-glucopyranose residues of the straight-chain glucan by introducing water molecules (**Figure 1A**). These enzymes comprise the superfamilies' GHs- 5–9, 12, 44, 45, 51, 74, and 124 (Lombard et al., 2014). The acid/base ($A^+/B^-$) catalytic mechanism of these enzymes to liberate mono- or oligo-saccharides, is facilitated by region as in endo- and exo-glucanses; and may exhibit chemical- (glucose, xylose, mannose, galactose, arabinose), and conformal-bias (retaining, inverting; **Figure 1A** and **Table 1**). Ancillary factors that contribute to substrate conversion include, the dominant secondary structural element (sse), presence and nature of the carbohydrate binding module(s), and the non-catalytic active site residues. The structural fold(s) for a retaining enzyme ($A^+/B^- =$ Glu/Glu) are $(\beta/\alpha)_8$ (GH5, 10, 26, 44, 51), and a β-jelly roll (GH7, 10). An inverting endoglucanase ($A^+/B^-$ = Asp/Asp or Asp/Glu), on the other hand, possesses the $(\alpha/\alpha)_6$ (GH6, 8, 9, 45, 48, 124) or seven-fold $\beta$-propeller (GH7) folds.

The carbohydrate binding module(s) (CBM) or cellulose binding domain(s) (CBD), present in these proteins dictate their association-dissociation kinetics with specific carbohydrate moieties and facilitate differential catalysis. These domains range from 40 (CBM1) to 200 amino acids (CBM17), and are present in several organisms (fungi, CBM1; bacteria, CBM2, 3, etc.; *D. discoideum*, CBM8; yeast, CBM54; plants, CBM49; Blume and Ennis, 1991; Koseki et al., 2008). The range of substrates bound include simple (galactose/lactose, CBM32; mannose, CBM13; chitin, CBM5, 12, 14, 18, 33; Newstead et al., 2005; Uni et al., 2009; Abramyan and Stajich, 2012; Li et al., 2015); compound (cellulose, CBM1-6, 8-11, 13, etc.; glycogen, CBM21, 48; starch, CBM20, 25; xylans, CBM22; polygalactouronic acid, CBM32; Abbott et al., 2007; Palomo et al., 2009; Janecek et al., 2011); and complex (lipopolysaccharide/lipoteichoic acid, CBM39; LacNAc; Bachman and McClay, 1996; Ficko-Blean and Boraston, 2006) molecules. Whilst, the *de facto* biological role for proteins with these domains is catalytic, there is ample evidence for the contrary (CBM1, 29, 43; Barral et al., 2005; Yoshida et al., 2005; Obembe et al., 2007). The cellulosome, is a complex of enzymes that participates in the assembly-disassembly of cellulose, along with several critical ancillary molecules (primer), and essential co-factors (Peng et al., 2002; Mansoori et al., 2014). This functional structure, along with molecular networks of co-expressed glycoside hydrolases is driven by the non-specific interactions of CBMs, either alone or in tandem with other enzyme-specific domains (Peng et al., 2002; Sharma et al., 2013; Mansoori et al., 2014). Perhaps the most intriguing constraint imposed by these domains is that of differential catalysis, i.e., same substrate, variable regions, and different enzymes. Naturally occurring cellulose is composed of at least two mutually exclusive regions: crystalline and amorphous (**Figure 1**). The inter- and intra-strand network of hydrogen bonds renders these microcrystalline regions well ordered, a feature that imposes an upper bound on the binding capacity of endoglucanases (CBM9, 49). In contrast, the latter, lacks this organization, permitting a far greater number of enzyme binding sites (CBM4, 6, 17, 28; Boraston et al., 2003; Jamal et al., 2004; Alahuhta et al., 2010). There are several hypotheses over the role of these domains in catalysis. These include, a physical, fix-and-stretch mechanism of the carbohydrate moiety from its parent glucan. This notion is based on the abundance of aromatic amino acids (W/F/Y) in these modules (Simpson et al., 2000; Roske et al., 2004; Flint et al., 2005; Tunnicliffe et al., 2005), and the presence of calcium (CBM35, 36, 60; Montanier et al., 2010).

Plant GH9 endoglucanases, like other members possess an activity profile that includes endoglucanase (EC3.2.1.4), lichenase (EC3.2.1.6, CBM4, 6, 13, 32, 54), mixed endoglucanase (EC3.2.1.73), exoglucanase (EC3.2.1.74, CBM2, 6, 10), cellobiohydrolase (EC3.2.1.91, CBM1-5, 10), and endo-xyloglucanase (EC3.2.1.151, CBM1-3, 30, 35, 44; **Figure 1B**) activity. The associated CBMs however, are not, always present together in a single sequence (Lombard et al., 2014). Extant classification schema into classes A, B, and C, are centered on sequence similarity, codon usage, distribution of intron-exon boundaries, and presence/absence of a trans-membrane domain (TM, class A) or secretory peptide (SP, class B; Mølhøj et al., 2002; Libertini et al., 2004). Whilst, the association between function and these indices is reasonably predictive, the similarity between sequences of classes B and C could potentially vitiate simpler clustering protocols. A sequence was identified in the class B cellulase subfamily (SlCel9C1; *S. lycopersicum*) which possessed a novel domain that was designated as CBM49 (IPR019028, PF09478; Urbanowicz et al., 2007a). This module is similar to the CBM2 in *C. filmi*, and implies, that this subset of enzymes could potentially function as a general purpose plant cellulase with catalysis of both, ordered and amorphous regions (class C activity; McLean et al., 2000; Boraston et al., 2004; Zhang et al., 2015). However, *in vitro* and *in vivo* experiments by these and other investigators, with the mature protein, suggest that this domain is removed prior to mature transcript formation, accounting for the refractoriness of this class to crystalline cellulose as a substrate *in vitro* (Urbanowicz et al., 2007a).

The role of GH9 enzymes in regulating plant physiology is unequivocal. The presence of the TM-domain localizes class A GH9 endoglucanases to the cell wall (primary, secondary), and associated structures (cell plate), thereby, dictating assembly by *de novo* cellulose biosynthesis in these regions (Nicol et al., 1998; Zuo et al., 2000; Sato et al., 2001; Mølhøj et al., 2002; Mansoori

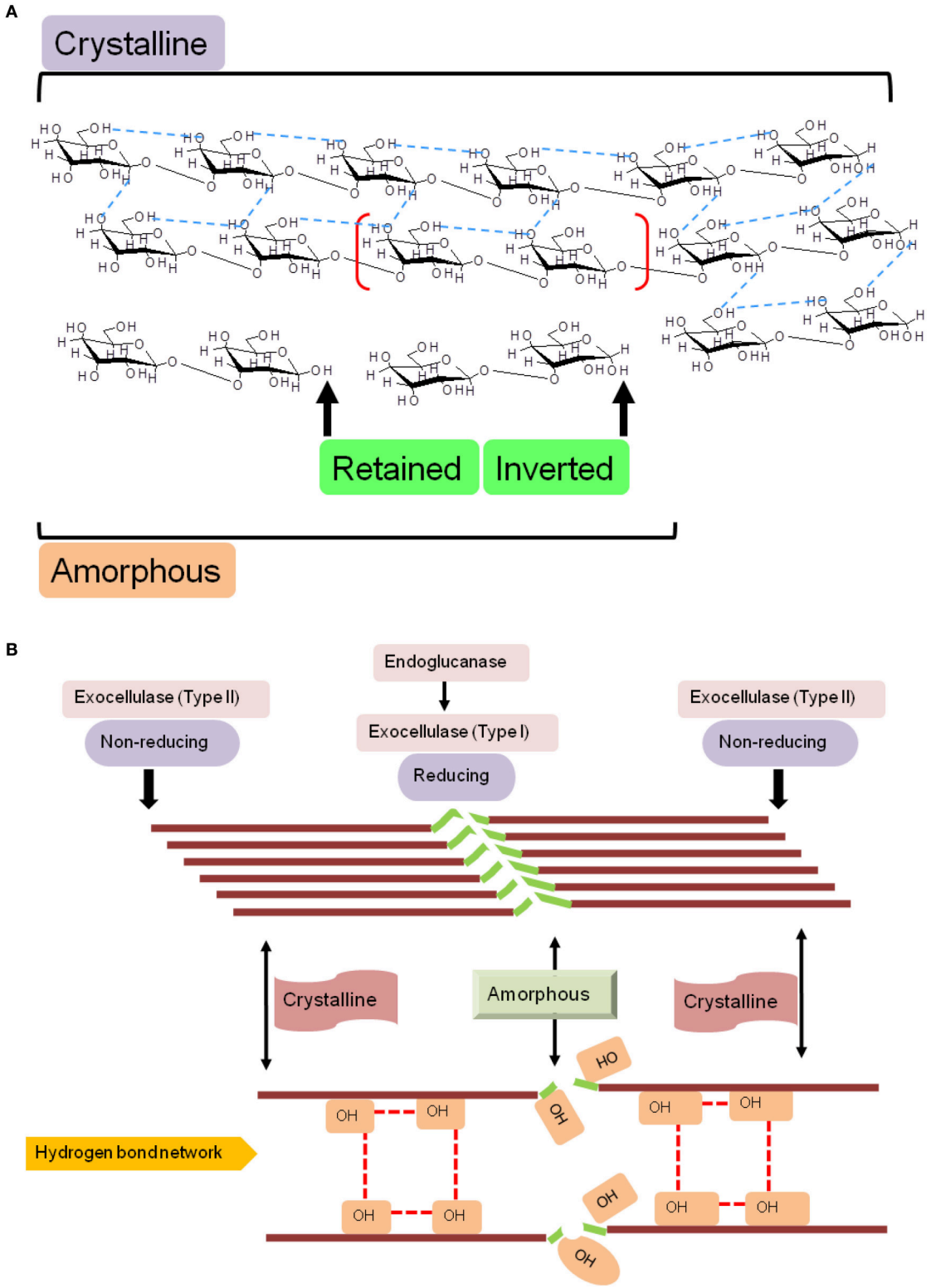



**FIGURE 1 | Schematic diagram of the molecular structure of cellulose. (A)** Naturally occurring cellulose is a straight chain glucan, and is a mixture of crystalline and amorphous regions, a nomenclature based on the proportion of intermolecular (inter- and intra-strand) hydrogen bonds. Cellulases, may retain or invert the hydroxyl ($OH^-$) on the anomeric carbon, a characteristic that is fold dependent, and **(B)** Structural and functional correlation of determinants of cellulose hydrolysis. The lack of a rigid structure (amorphous) renders cellulose amenable to catalytic cleavage by endoglucanase activity. Exocellulases (terminal), may act exclusively, but, usually effect digestion in association with endocellulases.

**TABLE 1 | Characteristics of GH9 endoglucanases associated with cellulase activity.**

| CAZy Family | Enzymatic activity (EC 3.2.1.x; EC 2.4.1.y) | Region (T/NT/TR/TN) | Mechanism (R/I) | Active site ($A^+/B^-$) | SSE |
|---|---|---|---|---|---|
| 5 | $x \in \{4, 8, 21, 25, 45, 58, 73, 74, 75, 78, 91, 104, 123, 132, 149, 151, 164, 168\}$ | NT | R | (E/E) | $(\beta/\alpha)_8$ |
| 6 | $x \in \{4, 91\}$ | TN | I | (D/D) | U |
| 7 | $x \in \{4, 73, 132, 176\}$ | TR | R | (E/E) | $2\,(\beta_8)$ |
| 8 | $x \in \{4, 8, 73, 132, 156\}$ | T;NT | I | (D/E) | $(\alpha/\alpha)_6$ |
| 9 | $x \in \{4, 6, 21, 73, 74, 91, 151, 165\}$ | NT | I | (D/E) | $(\alpha/\alpha)_6$ |
| 12 | $x \in \{4, 73, 151\}$ , $y \in \{207\}$ | NT | R | (E/E) | $2\,(\beta_8)$ |
| 44 | $x \in \{4, 151\}$ | NT | R | (E/E) | $(\beta/\alpha)_8$ |
| 45 | $x \in \{4\}$ | NT | I | (D/D) | $(\alpha/\alpha)_6$ |
| 48 | $x \in \{4, 14, 176\}$ | T; NT | I | (U/E) | $(\alpha/\alpha)_6$ |
| 51 | $x \in \{4, 8, 37, 55\}$ | NT | R | (E/E) | $(\beta/\alpha)_8$ |
| 74 | $x \in \{4, 150, 151\}$ | TR | I | (D/D) | $7\,(\beta_4)$ |
| 124 | $x \in \{4\}$ | NT | I | (U/U) | U |

*T, terminal; TR, terminal reducing (exo-); TN, terminal non-reducing (exo-); NT, non-terminal (endo-); R, Glucosyl retaining; I, Glucosyl inverting; A/B, residues for Acid/Base catalysis; D, Aspartic acid; E, Glutamic acid; U, Uncharacterized; SSE, secondary structural element; sfm, superfamily; SSEs: (β/α)8, 8-fold TIM (triosephosphate isomerase) barrel*; *(α/α)6, 6-fold barrel (Toroid)*; *7(β)4, 7-fold β-propeller*; *2(β)8, β-jelly roll.*

et al., 2014; Yu et al., 2014). Similarly, extracellular secretion, suggests a distributed influence and may facilitate a rapid response to stress (abiotic, biotic) by classes B and C enzymes. Class C GH9 endoglucanses, too, can influence development and response to stress), modify biofilm development for symbiotic or bacterial interactions, and can facilitate direct biomass conversion. Whilst, the high proportion of crystalline cellulose in the primary cell wall can be effectively and rapidly hydrolyzed (Urbanowicz et al., 2007a); its absence in the cell walls of root hair cells and endosperm, has also been attributed to active inhibition by this class of enzymes (AtGH9C1; *A. thaliana*; Shpigel et al., 1998; Sturcova et al., 2004; Otegui, 2007; del Campillo et al., 2012).

Complex polysaccharides, involving cellulose, are critical for host-bacterial interactions, and are secreted by the infecting bacteria, or activated in the host (plant roots/root hair, intestinal and lung epithelia; Cannon and Anderson, 1991; Mathee et al., 1999; Zogaj et al., 2001; White et al., 2003). The biofilms, thus formed facilitate aggregation, permit intercellular transfer of critical nutrients and signaling molecules, and can confer additional features (antibiotic resistance) by genetic exchange. A dual role for class C GH9 endoglucanases has been postulated and experimentally demonstrated in: (a) assisting infection, formation, and release into legume nodules by *Rhizobium* spp. (CelC2; Robledo et al., 2012), and (b) colonization by *Rhizobium* spp. and *A. tumefaciens*, by maturation/branching of this extracellular matrix (Matthysse et al., 1995; Robledo et al., 2012). The localization of AtGH9C1 in root hair cells (del Campillo et al., 2012), and concomitant infection with *A. tumefaciens* could increase the bacterial load (Matthysse et al., 2005), thereby, enhance the tumor forming capacity of these gram negative bacteria. Here, a combination of hydrolysis, translocation, and elongation-by-branching of cellulose by class C GH9 endoglucanases (plant, bacterial), would ensure optimal colonization. The most exciting role for class C GH9 endoglucanases, is their potential contribution to the biofuel industry (Lopez-Casado et al., 2008). Cellulose digestion can be mediated by the simultaneous presence of endo- and exo-glucanase regions in a single protein (CelA; *C. bescii*), a combinatorial association of endo- and exoglucanases in the cellulosome (CclEXL-1; *C. clariflavum*), and the possession of specialized modules (CBM49, CBM2) (Urbanowicz et al., 2007a; Chung et al., 2015; Artzi et al., 2016).

The methods and algorithms used to classify enzymes (superfamily, family, subfamily) depend on sequence based features or the conformational mapping of 3D information (secondary/tertiary/quaternary) to the primary structure. Supervised learning, is a machine learning method, that mandates training of an algorithm on well defined sets, and includes support vector machines (SVMs), regression analysis, neural networks, among several others. The SVM algorithm creates a hyperplane, and seeks to identify data points closest (support vectors) to this. It also entails an optimization to maximize the inter planar spacing. SVMs, for protein sequence classification will typically consider combinatorial associations of the amino acids sequence, such as pairs and triplets, etc. from the set of training sequences for feature extraction. HMMs, on the hand are models of a multiple sequence alignment, and represents a consensus of all the columns selected. SVMs, despite their predictive propensity, require unambiguous data and draw upon results from multiple rounds of pairwise comparisons (multiclass SVMs). Further, sequences with high sequence identity/similarity and common catalytic function (subfamily), might be better candidates for classification by SVM schema. GH9 endoglucanases have an intermediate level of sequence identity/similarity, with dominant function being clearly attributed to the presence of definitive region(s) in the N- (secretory peptide, transmembrane domain) and C-termini (CBM49) of the mature protein, and their linkage to the remainder of the protein (Mølhøj et al., 2002; Libertini et al., 2004; Urbanowicz et al., 2007a). The aforementioned enzyme specific constraints, and a superior performance assessment of

HMMs over SVMs, lends credence to our choice of HMM-ANN as the analytic platform to stratify GH9 endoglucanases (Khater and Mohanty, 2015).

There are a number of general Hidden Markov Model based predictors of protein function and classification (Gene3D, Pfam; Sonnhammer et al., 1998; Lees et al., 2010). These methods, despite providing initial pointers to novel candidate sequences, are unable to segregate closely related proteins. This limitation may be compensated, in-part, by populating the training dataset with sequences that meet stringent criteria, such as the availability of empirical data (Kundu, 2012). Alternate possibilities include, the use of pre-defined thresholds for data output, methods to screen the HMM output, and defining numerical patterns of domain dominance. In this study, we use a reverse look-up strategy to infer plant cellulose hydrolysis activity of putative sequences, from proteins which have been previously characterized. Since, our objective was to scan this data for highly probable class C sequences, a mathematical filter was developed to screen these on the basis of the HMM scores of the included profiles. Rigor of the prediction schema was ensured by formulating and validating indices to ascertain functionality from the returned results. Since the rules governing this association are complex, an ANN based-clustering protocol was chosen to infer and later, predict class assignment. The product of ANN-predicted values (weights, modifiers, constants) with one or more variables may be used to approximate a given function. This subset of supervised machine learning methods is non-linear and mandates the presence of high-confidence training datasets, but is able to delineate novel patterns with reasonable accuracy, and is suitably robust.

# METHODS

## Dataset Creation

We downloaded GH9 endoglucanse protein sequences from the CAZy (http://www.cazy.org) and UniProtKB (http://uniprot.org) databases (**Figure 2**). All associated information such as domain architecture, presence of specialized motifs, reaction chemistry, and physiological roles were cross referenced with InterPro (https://www.ebi.ac.uk/interpro), Pfam (http://pfam.xfam.org), SMART (http://smart.embl-heidelberg.de), and PROSITE (http://prosite.expasy.org) databases. Only endoglucanases with (a) demonstrable cellulase activity (EC 3.2.1.x), (b) transcript data with biochemical and/or physiological function, and (c)

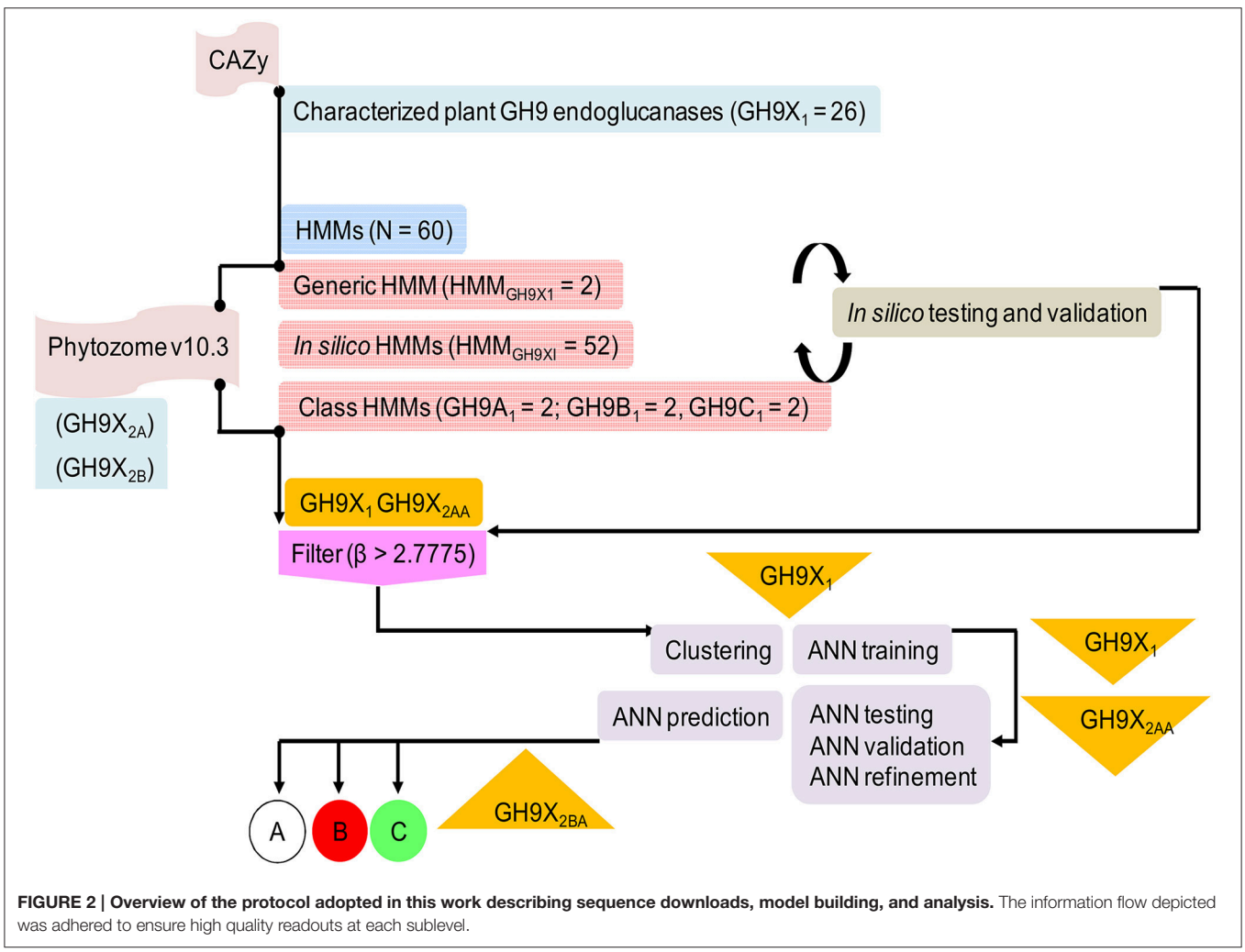



**FIGURE 2 | Overview of the protocol adopted in this work describing sequence downloads, model building, and analysis.** The information flow depicted was adhered to ensure high quality readouts at each sublevel.

an available 3D structure, were short listed. These sequences were clustered in accordance with previously defined criteria into classes A, B, and C cellulases (Sonnhammer et al., 1998; Apweiler et al., 2000; Letunic et al., 2002; Sigrist et al., 2010; Lombard et al., 2014). The 3D models of the submitted sequences were received from the Phyre2 server (Kelley et al., 2015). This preliminary compilation of 157 sequences was filtered, to exclude redundant data, such that the final dataset of GH9 sequences ($GH9X_1 = 26 = \{GH9A, GH9B, GH9C\}$) spanned 11 distinct plant genera (Table S1). Further partitioning, was based on the availability of enzyme specific experimental data ($GH9A_1 = 6$; $GH9B_1 = 16$; $GH9C_1 = 4$; Table S1). These sequences were used to construct the HMMs and train the ANN (**Table 2** and Table S2). Similarly, the sequences used to test these methods were divided into two groups: $GH9X_{2A} = 147$ (available expression data), and $GH9X_{2B} = 874$ (curated primary transcript data from the eukaryotic subclade Viridiplantae). The protein sequences for these datasets were downloaded from the Phytozome v10.3 (http://phytozome.jgi.doe.gov/pz/portal.html) database, and were mutually exclusive ($GH9X_1 \cap GH9X_2 = 0$). $GH9X_{2A}$, comprised sequences from *Arabidopsis thaliana* ($GH9X_{2A\_atha} = 24$), *Glycine max* $\left(GH9X_{2A\_gmax} = 39\right)$, *Oryza sativa* ($GH9X_{2A\_osat} = 25$), *Solanum lycopersicum* ($GH9X_{2A\_slyc} = 21$), and *Zea mays* ($GH9X_{2A\_zmay} = 38$) (Table S3A). Since, the sequences in the training set were few, $GH9X_{2A}$, was used to refine and partially validate the HMM and ANN predictions, as well as, corroborate their function *in vivo* (Zimmermann et al., 2004; Rensink et al., 2005; Skibbe et al., 2006; Cao et al., 2012).

## Construction of a Profile Database

We used the sequences of the training dataset ($GH9X_1$) to construct the HMMs ($HMM_{GH9X} = 60$) (**Table 2**; Text S1–S3), and broadly segregated into sequence- ($HMM_{1D} = 30$) and their corresponding structure-based ($HMM_{3D} = 30$) profiles. Since, a 3D- alignment is based on the conformational arrangement of secondary structural elements and active site residues, a higher correlation to function, of the corresponding HMM was subsumed. Alignments and cladograms for each dataset were generated separately with the STRAP (Structural Alignment of Proteins; http://www.bioinformatics.org/strap) and Clustal Omega v1.2.1suite of programs, format conversion was server-based (http://www.ibi.vu.nl/programs/convertalignwww), and HMMER 3.0 (http://hmmer.janelia.org) was used for model building, analysis, database construction, and similarity studies with the input sequences (Gille et al., 2014).

The highest scoring region/subdomain, for each HMM profile, of a sequence was considered for analysis (**Table 2**).

The large number of profiles (30 pairs = 1D and 3D) were utilized to: (a) query the putative proteome of 34 green plants and algae present in Phytozome v10.3, for putative GH9 endoglucanase homologs ($HMM_{GH9X}$; 1 pair), (b) ascertain the profile decomposition of each test sequence (3 pairs; Enzyme activity of sequence : $= \overline{HMM}_{GH9A}, \overline{HMM}_{GH9B}, \overline{HMM}_{GH9C}$), and (c) compensate, for the reduced size of the training sequences, by deploying an exhaustive leave-out-one strategy to compute, analyze, and computationally validate, the profile

**TABLE 2 | Summary and select details of HMM profiles utilized in this work.**

| | Profile | Sequences | | AL | ML | ES | RP |
|---|---|---|---|---|---|---|---|
| $HMM_{GH9X} = 2$ | E5Tx1D | 26 | | 754 | 520 | 0.8 | 0.592 |
| | E5Tx3D | | | 825 | 449 | 0.79 | 0.589 |
| $HMM_{GH9A} = 2$ | E4TA1D | 6 | | 622 | 618 | 0.42 | 0.592 |
| | E4TA3D | | | 615 | 495 | 0.5 | 0.592 |
| $HMM_{GH9B} = 2$ | E4TB1D | 16 | | 534 | 501 | 0.64 | 0.589 |
| | E4TB3D | | | 574 | 434 | 0.6 | 0.592 |
| $HMM_{GH9C} = 2$ | E4TC1D | 4 | | 651 | 625 | 0.48 | 0.589 |
| | E4TC3D | | | 622 | 566 | 0.53 | 0.592 |
| | E5Tx3D | | | 825 | 449 | 0.79 | 0.589 |
| | | **Training** | **Validation (UID)** | | | | |
| $HMM_{GH9AI} = 12$ | TAE11D | 5 | 1 (Q6X680) | 622 | 618 | 0.42 | 0.594 |
| | TAE13D | | | 600 | 511 | 0.48 | 0.588 |
| | TAE21D | 5 | 1 (O04890) | 622 | 619 | 0.42 | 0.596 |
| | TAE23D | | | 613 | 503 | 0.49 | 0.588 |
| | TAE31D | 5 | 1 (G0ZTA3) | 622 | 619 | 0.41 | 0.587 |
| | TAE33D | | | 615 | 490 | 0.47 | 0.587 |
| | TAE41D | 5 | 1 (Q6DMM4) | 622 | 618 | 0.42 | 0.591 |
| | TAE43D | | | 606 | 499 | 0.48 | 0.594 |
| | TAE51D | 5 | 1 (D3JWK8) | 622 | 618 | 0.42 | 0.592 |
| | TAE53D | | | 619 | 493 | 0.49 | 0.591 |
| | TAE61D | 5 | 1 (Q38890) | 619 | 619 | 0.42 | 0.588 |
| | TAE63D | | | 590 | 511 | 0.47 | 0.585 |
| $HMM_{GH9BI} = 32$ | TBE11D | 15 | 1 (Q42872) | 550 | 490 | 0.62 | 0.589 |
| | TBE13D | | | 529 | 437 | 0.6 | 0.589 |
| | TBE21D | 15 | 1 (O04972) | 556 | 490 | 0.62 | 0.588 |
| | TBE23D | | | 554 | 442 | 0.6 | 0.593 |
| | TBE31D | 15 | 1 (Q41012) | 546 | 490 | 0.62 | 0.59 |
| | TBE33D | | | 554 | 437 | 0.6 | 0.592 |
| | TBE41D | 15 | 1 (Q93WZ0) | 547 | 492 | 0.63 | 0.592 |
| | TBE43D | | | 554 | 437 | 0.6 | 0.589 |
| | TBE51D | 15 | 1 (Q93WZ1) | 539 | 494 | 0.63 | 0.592 |
| | TBE53D | | | 560 | 437 | 0.6 | 0.591 |
| | TBE61D | 15 | 1 (P22503) | 537 | 496 | 0.61 | 0.588 |
| | TBE63D | | | 552 | 437 | 0.59 | 0.593 |
| | TBE71D | 15 | 1 (Q40763) | 556 | 490 | 0.62 | 0.589 |
| | TBE73D | | | 552 | 434 | 0.6 | 0.592 |
| | TBE81D | 15 | 1 (Q9XIY8) | 537 | 495 | 0.63 | 0.589 |
| | TBE83D | | | 566 | 435 | 0.6 | 0.591 |
| | TBE91D | 15 | 1 (Q6DMM3) | 553 | 490 | 0.62 | 0.591 |
| | TBE93D | | | 564 | 438 | 0.6 | 0.59 |

*(Continued)*

**TABLE 2 | Continued**

| | Profile | Sequences | | AL | ML | ES | RP |
|---|---|---|---|---|---|---|---|
| | TBE101D | 15 | 1 (P94114) | 537 | 494 | 0.63 | 0.592 |
| | TBE103D | | | 563 | 439 | 0.6 | 0.588 |
| | TBE111D | 15 | 1 (Q42875) | 534 | 490 | 0.62 | 0.59 |
| | TBE113D | | | 560 | 434 | 0.59 | 0.588 |
| | TBE121D | 15 | 1 (Q42871) | 528 | 489 | 0.61 | 0.591 |
| | TBE123D | | | 555 | 436 | 0.58 | 0.587 |
| | TBE131D | 15 | 1 (O82473) | 552 | 492 | 0.62 | 0.589 |
| | TBE133D | | | 556 | 440 | 0.6 | 0.587 |
| | TBE141D | 15 | 1 (Q9CAC1) | 544 | 490 | 0.62 | 0.587 |
| | TBE143D | | | 558 | 434 | 0.6 | 0.592 |
| | TBE151D | 15 | 1 (Q9SRX3) | 545 | 492 | 0.62 | 0.588 |
| | TBE153D | | | 555 | 437 | 0.6 | 0.589 |
| | TBE161D | 15 | 1 (Q9ZTL0) | 543 | 493 | 0.63 | 0.593 |
| | TBE163D | | | 560 | 434 | 0.59 | 0.587 |
| $HMM_{GH9CI} = 8$ | TCE11D | 3 | 1 (Q8LJP6) | 648 | 625 | 0.47 | 0.592 |
| | TCE13D | | | 609 | 579 | 0.5 | 0.588 |
| | TCE21D | 3 | 1 (Q5NAT0) | 636 | 621 | 0.43 | 0.588 |
| | TCE23D | | | 605 | 575 | 0.47 | 0.59 |
| | TCE31D | 3 | 1 (Q93WY9) | 647 | 621 | 0.47 | 0.589 |
| | TCE33D | | | 626 | 557 | 0.51 | 0.588 |
| | TCE41D | 3 | 1 (Q9ZSP9) | 642 | 625 | 0.47 | 0.59 |
| | TCE43D | | | 612 | 566 | 0.52 | 0.592 |

*UID, Uniprot Identifier; $HMM_{GH9X}$, Dominant function of sequence (X = A, B, C); Text S1 and S3 in Supplementary Material; $HMM_{GH9XI}$, Dominant function of sequence for in silico abalysis (X = A, B, C); Text S2 in Supplementary Material; AL, Alignment Length; ML, Model Length; ES, Effective number of Sequences; RP, Relative Entropy per Position.*

HMM scores for each training sequence $HMM_{GH9XI} = 26\ \text{pairs} = HMM_{GH9AI} + HMM_{GH9BI} + HMM_{GH9CI}$) (**Figure 2**, **Table 2**; Table S2A). Here, every sequence was assumed, *a priori*, to possess dual membership, i.e., it was part of both the training and validation subsets for a particular profile. The raw HMM scores of the selected profiles of only those sequences that were used for validation, were considered and averaged (1D, 3D) This can be generalized as:

$$GH9XI = \begin{cases} \left(HMM_{GH9AI} - ((2)\,(GH9A_1 - 1))\right), \\ \quad (HMM_{GH9BI}), (HMM_{GH9CI}),\ X = A \qquad Def.1 \\ (HMM_{GH9AI}),\ \left(HMM_{GH9BI} - ((2)\,(GH9B_1 - 1))\right), \\ \quad (HMM_{GH9CI}),\ X = B \qquad Def.2 \\ (HMM_{GH9AI}),\ (HMM_{GH9BI}), (HMM_{GH9CI} - ((2) \\ \quad (GH9C_1 - 1))),\ X = C \qquad Def.3 \end{cases}$$

Consider the following example. Since, the number of training sequences with class C activity are only four ($GH9C_1 = 4$; number of class C profiles = 8), in the leave-one-out schema, only data from this single class C sequence (number of class C profiles = 2) was deemed relevant. Similarly, for this particular sequence all class A and B profiles scores would be taken into account (class A profiles = 12; class B profiles = 32). Thus, the combined HMM scores of these relevant profiles were considered ($12 + 32 + 2 = 46$ or 23 pairs), for this class C sequence (**Table 2**, Tables S2A,B).

## Screening Filter

Profile HMMs (pHMMs), whilst being theoretically well grounded, do not offer unambiguous predictions, i.e., the query sequence is a function of the included profiles. Although, the resultant data may be filtered with the use of inclusion and threshold scores, an inter-profile comparison of scores with defined exclusion criteria is clearly desirable. Populating the ANN input with sequences with well-spaced HMM scores, can be accomplished by progressively screening out sequences which do not comply with these conditions. This filter compares the raw HMM scores of the constituent profiles, and outputs a quality score ($\beta$; Equation 1). The method is based on computing a modified Z-score of pairs of groups of profiles that comprise a sequence, i.e., a mixture of classes A, B, and C, calculated as $\left(C\,(3,2) = \binom{3}{2} = 3\right)$. Here,

$$\begin{aligned} Group12: &= \left\{\left(\overline{HMM}_{GH9AI}, \overline{HMM}_{GH9BI}\right), \right. \\ &\quad \left.\left(\overline{HMM}_{GH9BI}, \overline{HMM}_{GH9CI}\right)\right\} \qquad Def.4 \\ Group23: &= \{(\overline{HMM}_{GH9BI}, \overline{HMM}_{GH9CI}), \\ &\quad (\overline{HMM}_{GH9AI}, \overline{HMM}_{GH9CI})\} \qquad Def.5 \\ Group13: &= \{(\overline{HMM}_{GH9AI}, \overline{HMM}_{GH9BI}), \\ &\quad (\overline{HMM}_{GH9AI}, \overline{HMM}_{GH9CI})\} \qquad Def.6 \\ \beta &= 1/2\,(\sum_{i=1}^{i=3}\sum_{j=1}^{j=3}\alpha_{ij})\ \forall\ i \neq j,\ \alpha_{ij} = \alpha_{ji} \qquad (1) \\ \alpha_{ij} &= gp_{ij} = \left(|\mu_i - \mu_j|/100\right) \\ &\quad \left(|\mu_i - \mu_j|/\left(\sqrt[2]{\left(\sigma_i^2 + \sigma_j^2/\tau\right)}\right)\right) \qquad (2) \\ i, j &\in \{1, 2, 3\} with\ i \neq j \\ \tau &= 2\ (members\ in\ each\ group) \\ \mu &:= intra - group\ mean \\ \sigma^2 &:= intra - group\ variance \\ \mathrm{Z} &:= inter - group\ z - score \end{aligned}$$

The final selection of the threshold value was a numerical refinement of $\min(\beta)$ of $GH9X_1$ on $GH9X_{2A}$ (**Figures 4A,B**), and its subsumed correspondence with the inter profile HMM difference ($\min(\beta) \mapsto median(\Delta HMM)$; *Equation* 8) (Tables S2C, S3C).

## ANN Based Assignment of Dominant Enzymatic Activity of GH9 Endoglucanases

As highlighted *vide supra*, confirmation of enzymatic activity can be unequivocally resolved only in a laboratory setting. Models, at best, offer the probability of a particular outcome. This measure of uncertainty is compounded by pHMM-based analytics and the paucity of experimental data. The native profile HMM scores *(P)* of a query sequence is one measure of ascertaining the function

of a putative protein, i.e., max *(P)*. However, the proximity of these scores, especially in classes B and C sequences, precludes confidence in any such assignment. Descriptive statistics of these pairs-of-pairs means ($\alpha_{12}$, Group 12; $\alpha_{23}$, Group 23; $\alpha_{13}$, Group 13), suggests that this modification of the Z-score (Equations 2, 6-8), may provide a rigorous framework that could not only exclude sequences with equivalent profile HMM scores, while at the same time be used ($\beta$; Equation 1) to cluster sequences.

Cluster analysis (k-means), of the $\beta$-values of each sequence of $GH9X_1$, was used to compute class-specific cluster means ($k = 3$; $\beta'_A$, $\beta'_B$, $\beta'_C$) which were then graphed and plotted using a cluster-dendrogram (**Figures 4C,D**, Text S4). This value was chosen so as to maximize the distance between the centroids of the clusters, thereby, ensuring high confidence in the assignments ($\max\left(between_{SS}/total_{SS}\right)$). Outliers, were removed to ensure rigor (**Figure 4D**). These $\beta'$- values were assumed to be linear combination of the weighted derived scores for each sub-group ($\beta' \cong \sum_{i\,=\,1,2;j\,=\,2,3}(\gamma_{ij})(\alpha_{ij}), \forall\, i \neq j$) of a particular sequence. The values of these weights (Text S5), and their confidence at 0.95 ($1 - \alpha|\alpha = 0.05$) were computed using an ANN (Hidden layers = 10; threshold = 0.01). The predicted ANN values ($\beta''$) in this leave-one-out ($GH9X_{1A} = 24$) approach, were then compared with the previously computed cluster means ($\beta'' \cong \beta'$).

The absence of confirmatory enzyme kinetic data for the test sequences, i.e., mRNA ($GH9X_{2A}$) and genomic-hypothetical proteins ($GH9X_{2B}$), precludes the direct usage of cluster means ($\beta'$) or their approximations ($\beta''$), as unambiguous predictors of enzyme function. Here, instead, it was reasoned that an enzyme specific class interval, rather than a single value, for the HMM-ANN prediction on $GH9X_{2A}$ ($\min\left(\beta''_{GH9A}\right) \leq \beta''_{GH9A} \leq \max\left(\beta''_{GH9A}\right)$; $\min\left(\beta''_{GH9B}\right) \leq \beta''_{GH9B} \leq \max\left(\beta''_{GH9B}\right)$; $\min\left(\beta''_{GH9C}\right) \leq \beta''_{GH9C} \leq \max\left(\beta''_{GH9C}\right)$) for each enzyme function, along with select patterns of the computed $\alpha_{ij}$-values, may encompass function more effectively. The R-scripts (R-3.0.0) needed to analyze this data, and perform other miscellaneous tasks were coded in-house, or downloaded as packages. This included the ANNs (neuralnet), clustering, and plotting (cluster; fpc). Chemical structures were drawn using the ChemSketch suite (freeware) installed locally.

## Validating the Integrated Pipeline

The exhaustive leave-one-out strategy utilized for the computations, also functioned to cross validate (LOOCV) the predictions by the HMMs and the ANN, and was chosen to compensate for the paucity of training sequences. The criteria to validate, for selecting the appropriate HMM, was the equivalence of the highest scoring profile of a sequence with predicted enzyme function (Enzyme activity of a sequence = $\max\left(\overline{HMM}_{GH9A}, \overline{HMM}_{GH9B}, \overline{HMM}_{GH9C}\right)$)., as a generalization for the analytic and *in silico* steps, as under:

$$GH9A := \overline{HMM}_{GH9A} > \left\{\overline{HMM}_{GH9B}, \overline{HMM}_{GH9C}\right\} \quad Def.7$$

$$GH9B := \overline{HMM}_{GH9B} > \left\{\overline{HMM}_{GH9A}, \overline{HMM}_{GH9C}\right\} \quad Def.8$$

$$GH9C := \overline{HMM}_{GH9C} > \left\{\overline{HMM}_{GH9A}, \overline{HMM}_{GH9B}\right\} \quad Def.9$$

$$GH9AI := \left(HMM_{GH9AI} - \left((2)\left(GH9A_1 - 1\right)\right)\right) > \left\{\left(HMM_{GH9BI}\right), \left(HMM_{GH9CI}\right)\right\} \quad Def.10$$

$$GH9BI := \left(HMM_{GH9BI} - \left((2)\left(GH9B_1 - 1\right)\right)\right) > \left\{\left(\overline{HMM}_{GH9AI}\right), \left(\overline{HMM}_{GH9CI}\right)\right\} \quad Def.11$$

$$GH9CI := \left(HMM_{GH9C_1} - \left((2)(GH9C_1 - 1)\right)\right) > \left\{\left(\overline{HMM}_{GH9AI}\right), \left(\overline{HMM}_{GH9BI}\right)\right\} \quad Def.12$$

Similarly, the index of measurement, chosen, to ascertain relevance of the ANN predicted values ($\beta''$) was based on the following:

$$\beta'' \cong \beta' \quad (3)$$

$$\because \beta' \equiv max\left(\overline{HMM}_{GH9A}, \overline{HMM}_{GH9B}, \overline{HMM}_{GH9C}\right) \quad (4)$$

$$\therefore \beta'' \equiv max\left(\overline{HMM}_{GH9A}, \overline{HMM}_{GH9B}, \overline{HMM}_{GH9C}\right) \quad (5)$$

The chi-squared ($\chi^2$) statistic was used to compare the two sets of numerical data points for $GH9X_1$ (Equation 3). Since, these values were based on a restricted dataset, the procedure was repeated on $GH9X_{2A}$. However, despite the availability of expression data for these sequences, information on the catalytic activity of their encoded proteins is undefined, and therefore, at best inferred (Equation 5).

## Analysis of Biological Significance of the ANN-Based Predictions Using Transcriptomic Data

The relevance of these predictions was assessed using available gene expression datasets. For *O. sativa* and *A. thaliana*, extensive expression data for the anatomical and developmental stages is publically available. These were analyzed to observe the fluctuations in gene expression of some of the sequences identified in dataset ($GH9X_{2A}$). The metadata for gene expression analysis in rice was downloaded from the rice oligonucleotide array database (ROAD; http://www.ricearray.org; Table S6A), whereas, the same for *A. thaliana* was extracted using GENEVESTIGATOR (https://genevestigator.com/gv; Table S6B; Zimmermann et al., 2004; Cao et al., 2012).

# RESULTS

## Salient Features of Models and *In silico* Analysis

HMMs, with sequence and domain threshold values ($E_{seq} = E_d = 10E-06$), were generated from the formatted alignments of amino-acid sequences and their corresponding modeled 3D-coordinates (**Figures 2**, **3**). Since, the phylogenetic clades were similar for both sets of profiles (**Figures 3A,B**), profile pairs were averaged:

$$HMM_{GH9A} := average\left(HMM_{GH9A1D}, HMM_{GH9A3D}\right) \quad (Def.13)$$

$$HMM_{GH9B} := average\left(HMM_{GH9B1D}, HMM_{GH9B3D}\right) \quad (Def.14)$$

$$HMM_{GH9C} := average\left(HMM_{GH9C1D}, HMM_{GH9C3D}\right) \quad (Def.15)$$

The model(s) characteristics and summary are presented (**Table 2**). A fraction of the data ($\delta_A = HMM_{GH9AI}/$

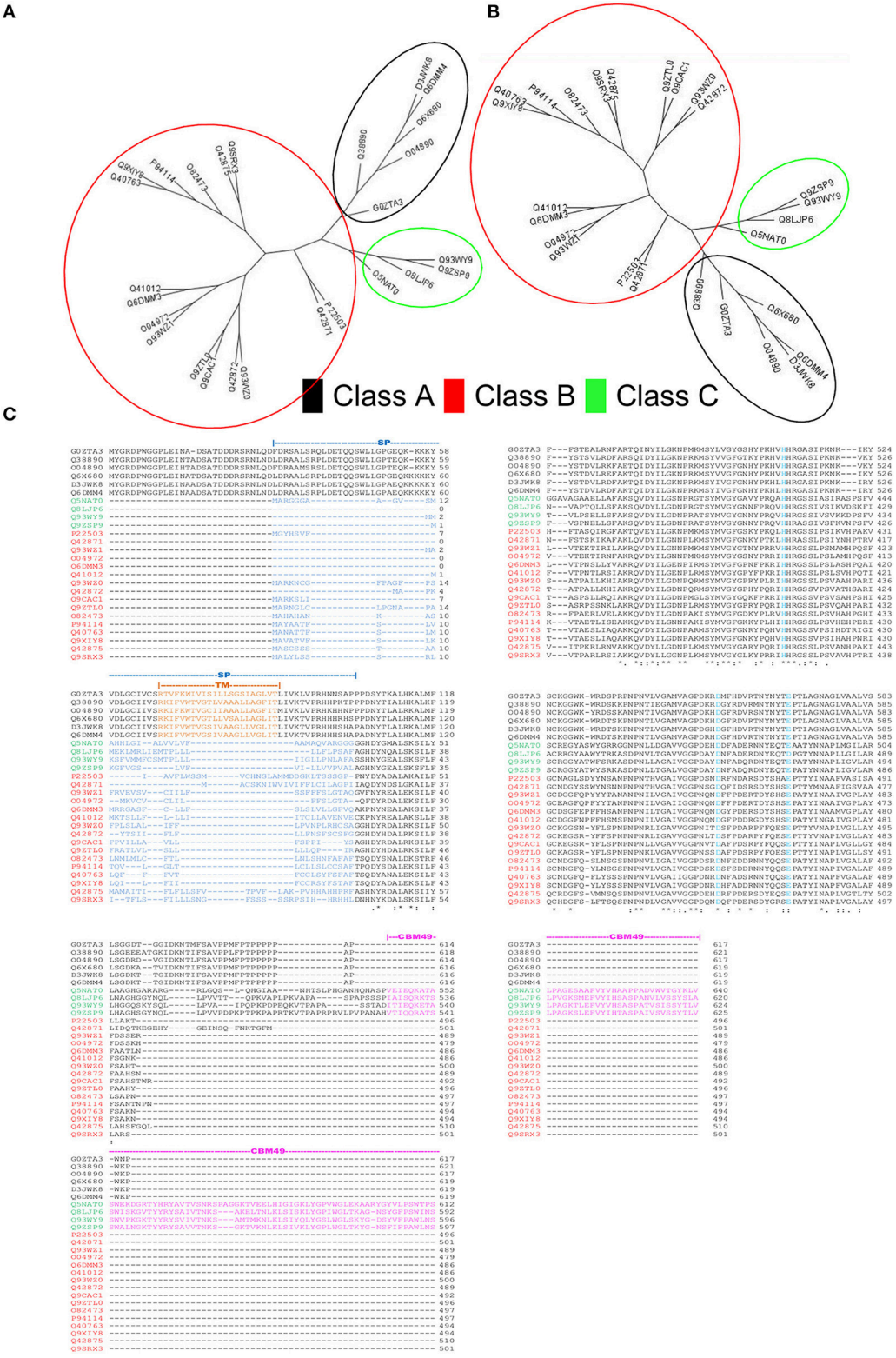



**FIGURE 3 | Rationale for computational strategies deployed in this work. (A,B)** Comparison of the unrooted cladograms of sequence- (1D) and structure- (3D) based alignments results in a congruent grouping of sequences, a key assumption in the computation of average pHMM-scores. **(C)** Multiple sequence alignment of the sequences present in the training set. Whilst, class A sequences are characterized by a well defined transmembrane domain (TM), classes B and C possess a signal peptide sequence at their N-terminal ends. The carbohydrate binding module 49 (CBM49), defines presence of class C activity, and is present at the C-terminal and is ≈100–120 AA.

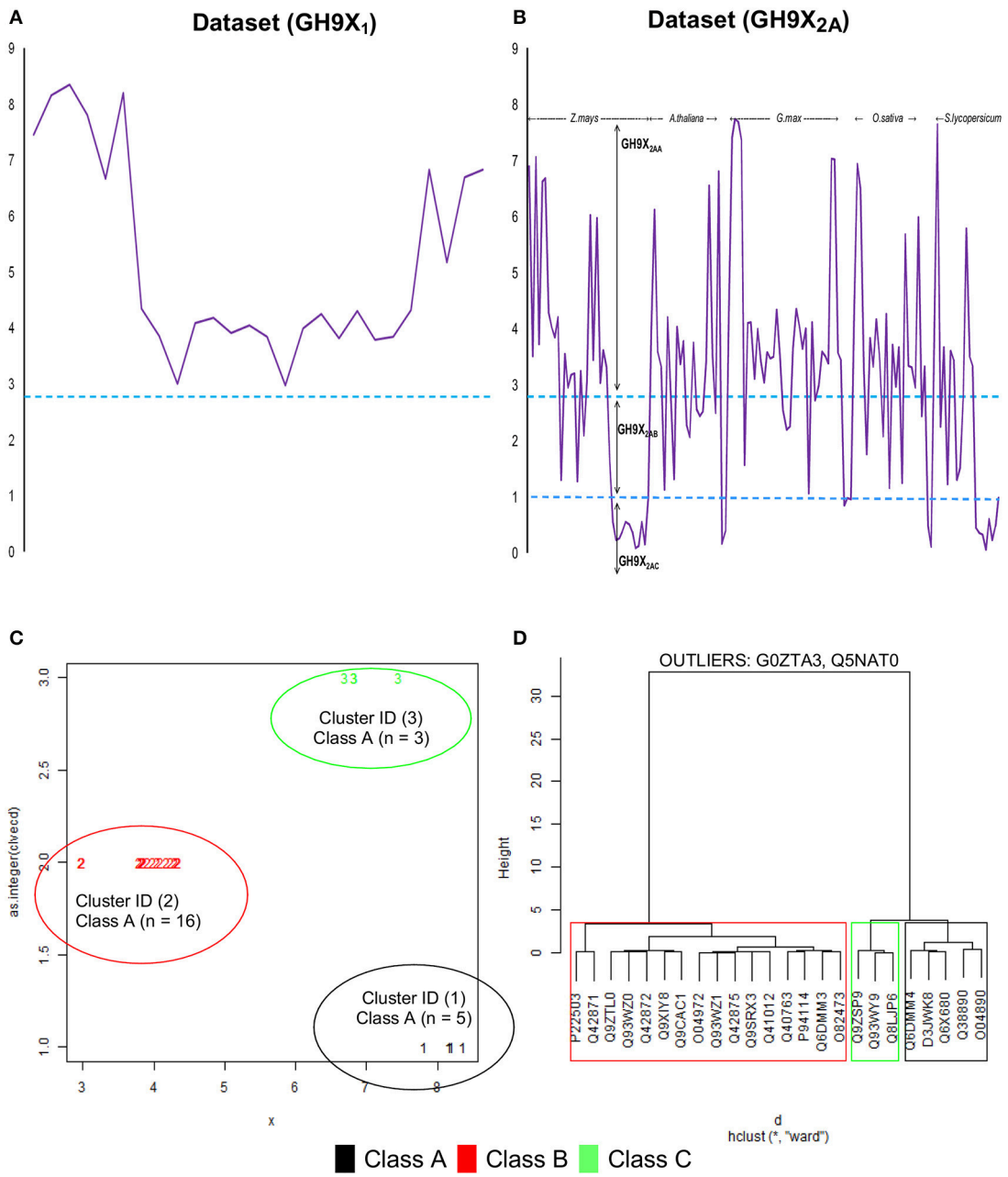



**FIGURE 4 | Computational analysis of datasets GH9X$_1$ and GH9X$_{2A}$. (A,B)** Trend of $\beta$-values across sequences in representative datasets. Here, $\beta = 2.777, 1.00$ and functions to identify sequences with well-spaced interprofile HMM scores ($GH9X_{2AA} = 92$; $GH9X_{2AB} = 28$; $GH9X_{2AC} = 27$) **(C)** k-means clustering of the centroids ($\beta$) of the clusters that correspond to the previously defined classes A, B, and C of plant GH9 endoglucanases after outlier exclusion ($GH9X_{1A} = 24$). **(D)** Cluster-dendrogram distribution of sequences of the modified training set ($GH9A_{1A} = 5$, $GH9B_{1A} = 16$, $GH9C_{1A} = 3$).

$HMM_{GH9XI}$; $\delta_B = HMM_{GH9BI}/HMM_{GH9XI}$; $\delta_C = HMM_{GH9CI}/HMM_{GH9XI}$) was used for computing the profile scores (Tables S2A,B). Thus, for any class A sequence ($HMM_{GH9AI} = 42, \delta_A \cong 0.72$; $Def.7$). The equivalent data for any class B ($HMM_{GH9BI} = 22, \delta_B \cong 0.38$; $Def.8$), and C $\left(HMM_{GH9CI} = 46, \delta_C \cong 0.79; Def.9\right)$ sequence was computed similarly (Tables S2A,B).

## Numerical Determination and Relevance of the Filter for Sequence Selection

The computation and selection of the $\beta$-values for the training sequences was based on the *in silico* selection of profiles ($HMM_{GH9XI}$) as outlined earlier. Here, $min(\beta) \cong 3.00$ (UID, P22503) (**Figure 4A** and Table S2C), corresponded to a median inter-profile HMM score $\cong$ 200. Refinement of

these computed values was accomplished by analyzing the fluctuations in $\beta$-values on $GH9X_{2A}$, and in the intervals ($[0, 50)$ , $[50, 150)$ , $[150, \infty)$) (**Table 3**). The sequence SOLYc05g052530.1.1 had a median interprofile HMM score $\cong$ 93, and a high $\beta$-value ($\beta_{Solyc05g052530.1.1} \cong 2.777$) (Tables S3B,C). Therefore, $(\beta > 2.777) \wedge \left(median\,(\triangle HMM) \geq 200\right)$, was chosen as the major criteria to further partition and evaluate sequences of $GH9X_{2A}$ ($GH9X_{2AA} = 92$, $GH9X_{2AB} = 28$, $GH9X_{2AC} = 27$), and (**Figure 4B**, **Table 3** and Table S3C). The proximity of inter profile HMM $\left(median\,(\triangle HMM)\right)$ scores for sequences with low $\beta$-values ($(0.67)\,(GH9X_{2AC}) \in [0, 50)$), whilst the ambiguity of these for sequences with intermediate $\beta$-values $\left((0.3–0.4)\,(GH9X_{2AB}\right) \in \{[0, 50)\,, [50, 150)\,, [150, \infty)\}$); along with their poor precision and recall precluded their selection as numerical approximations of the threshold value (**Table 3** and Table S3C). Conversely, the homogeneity of computed data $\left((0.86)\,(GH9X_{2AA}) \in [150, \infty)\right)$, perfect precision, and high recall, for sequences with $\beta > 2.777$, dictated the final value of the filter for shortlisting sequences ($GH9X_{2BA}$) (**Table 3**, Table S5).

We chose the subset ($GH9X_{2AA} = 92$), since these sequences had a high SNR/widely spaced interprofile HMM scores and therefore possessed adequate class specific discriminatory data ($GH9A_{2AA} = 19, GH9B_{2AA} = 43, GH9C_{2AA} = 30$). The data from these sequences was, used to a) refine the numerical value of $\beta$, and b) define the intervals for the computed ANN scores, used subsequently for unambiguous class assignment (**Figure 5**; Table S4). The analysis (precision and recall; **Table 3**) further, served as an index against overestimating the predictions by the ANN on sequences of $GH9X_{2BA}$ (**Table 3** and Table S5). The organism specific distribution was: *A. thaliana* ($GH9X_{2AA\_atha} = 14$), *G. max* $\left(GH9X_{2AA\_gmax} = 30\right)$ , *O. sativa* ($GH9X_{2AA\_osat} = 18$), *S. lycopersicum* ($GH9X_{2AA\_slyc} = 8$), and *Z. mays* ($GH9X_{2AA\_zmay} = 22$) (**Figure 4B**; Table S4). The putative GH9 endoglucanase homologs were identified ($HMM_{GH9X}$; $GH9X_{2B}$), downloaded fom Viridiplantae, filtered ($GH9X_{2BA} = 552$), and analyzed with the generic class specific HMMs ($HMM_{GH9A}$, $HMM_{GH9B}$, $HMM_{GH9C}$) (Table S5). Interestingly, only five sequences were excluded, despite a seven-order difference in magnitude ($logE_{GH9}/E_{GH9X}$) of the respective *E*-value thresholds.

## Predicted Activity of GH9 Endoglucanases

The clustered data ($\beta$; $GH9X_1$) (**Figures 4C,D**), was analyzed, wherein, a single cluster-node for each enzyme class was identified ($\beta'_{GH9A} \cong 8.134881$; $\beta'_{GH9B} \cong 3.912766$; $\beta'_{GH9} \cong 6.953259$; $between_{SS}/total_{SS} = 0.96$) (Text S4). Two outliers (UIDs G0ZTA3, Q5NAT0) (**Figure 4D**; Tables S2C,D) were identified, and excluded from further analysis. The class specific cluster means was approximated by the group scores of each training sequence ($\beta' \sim \alpha_{12} + \alpha_{23} + \alpha_{13}$; $GH9X_{1A} = 24$). The ANN scores $\left(\beta''\right)$, thus computed (leave-one-out) when compared with the cluster means $\left(\chi^2 = 0.005;\, df = 23;\, p = 0.001\right)$ (**Figure 5**; Table S2D), clearly suggest the equivalence of the ANN predictor with the dominant enzyme function of the sequence of interest (Equations 3–5; ANN prediction $\cong$ cluster mean of a class $\cong \max\left(\overline{HMM}_{GH9A}, \overline{HMM}_{GH9B}, \overline{HMM}_{GH9C}\right)$. The above criteria was used on $GH9X_{2AA}$ ($\chi^2 = 0$; $df = 91$) to associate, dominant enzyme function with statistically defined class specific intervals of the ANN-predictors ($\beta''_{GH9A}, \beta''_{GH9B}, \beta''_{GH9C}$) and $\alpha_{ij}$-values (**Table 4** and Table S4).

The bounds, thus defined, were then used to stratify the scores of the test sequences in $GH9X_{2BA}$ (**Tables 4**, **5** and Table S5). Our results suggest the following taxonomic distribution of GH9 proteins ($GH9A_{2BA} \approx 6\%$, $GH9B_{2BA} \approx 50\%$, $GH9C_{2BA} \approx 44\%$) (**Figures 6A,B**). Additional findings include $GH9B_{2BA} \cong GH9C_{2BA}$, $GH9A_{2BA} \cong GH9C_{2BA}$, $GH9C_{2BA} > GH9B_{2BA}$, $GH9B_{2BA} > (2)\,(GH9C_{2BA}) \cong GH9A_{2BA}$ (**Table 5** and **Figures 6C,D**). The following gene(s) of *O. sativa* (LOC_Os02g05744, LOC_Os09g23084, LOC_Os02g50490, LOC_Os08g32940) and *A. thaliana* (AT2G32990) were reannotated by our ANN-predictor as class C (Buchanan et al., 2012; Xie et al., 2013). The sequences, annotated by our prediction scheme ($GH9A_{2BA} = 31, GH9B^{+}{}_{2BA} = 231, GH9B^{-}{}_{2BA} = 49, GH9C_{2BA} = 241$) are available in fasta format (Text S6–S9).

## Assessment of Predictor Performance

The *in silico* HMM profiles for the training sequences when assessed, as defined (Defs.1–3, 10–12), using the LOOCV, had a precision of 100% ($GH9X_1 = 26$) (Tables S3A,B). The precision of the ANN computed approximations of the cluster means, and using the LOOCV, on the training set after outlier exclusion ($GH9X_{1A} = 24$) was 100% when assessed by the aforementioned criteria (Equation 5). Overestimation by our HMM-ANN pipeline was examined by the performance of these predictions on $GH9X_{2A}$ (**Table 3**).

## Meta-Analysis and Significance of the ANN-Based Prediction Schema

The predicted distribution of rice GH9 endoglucanases ($GH9A_{2AA-osat} = 3, GH9B_{2AA\_osat} = 8, GH9C_{2AA\_osat} = 7$) was examined across several tissues (Table S6). Most putative

**TABLE 3 | Performance of ANN-predictor on $GH9X_{2A}$.**

| Interval | $\beta$ | NSeq | M | MM | NM | P | R | △HMM (Observations) |
|---|---|---|---|---|---|---|---|---|
| [0, 50) | $\beta \geq 2.777$ | 92 | 92 | 0 | 0 | 100% | 85.2% | 224.475 (276) |
| [50, 150) | $1.00 \leq \beta < 2.777$ | 28 | 7 | 8 | 13 | 25% | 6.5% | 133.1 (84) |
| [150, ∞) | $0.00 < \beta < 1.00$ | 27 | 9 | 6 | 12 | 33% | 8.3% | 32.8 (81) |

*β, Value of statistical filter; Nseq, Number of sequences; M, Number of matches; MM, Number of mismatches; NM, Number of no matches; P, Precision; R, Recall; △HMM, Median HMM interprofile difference.*

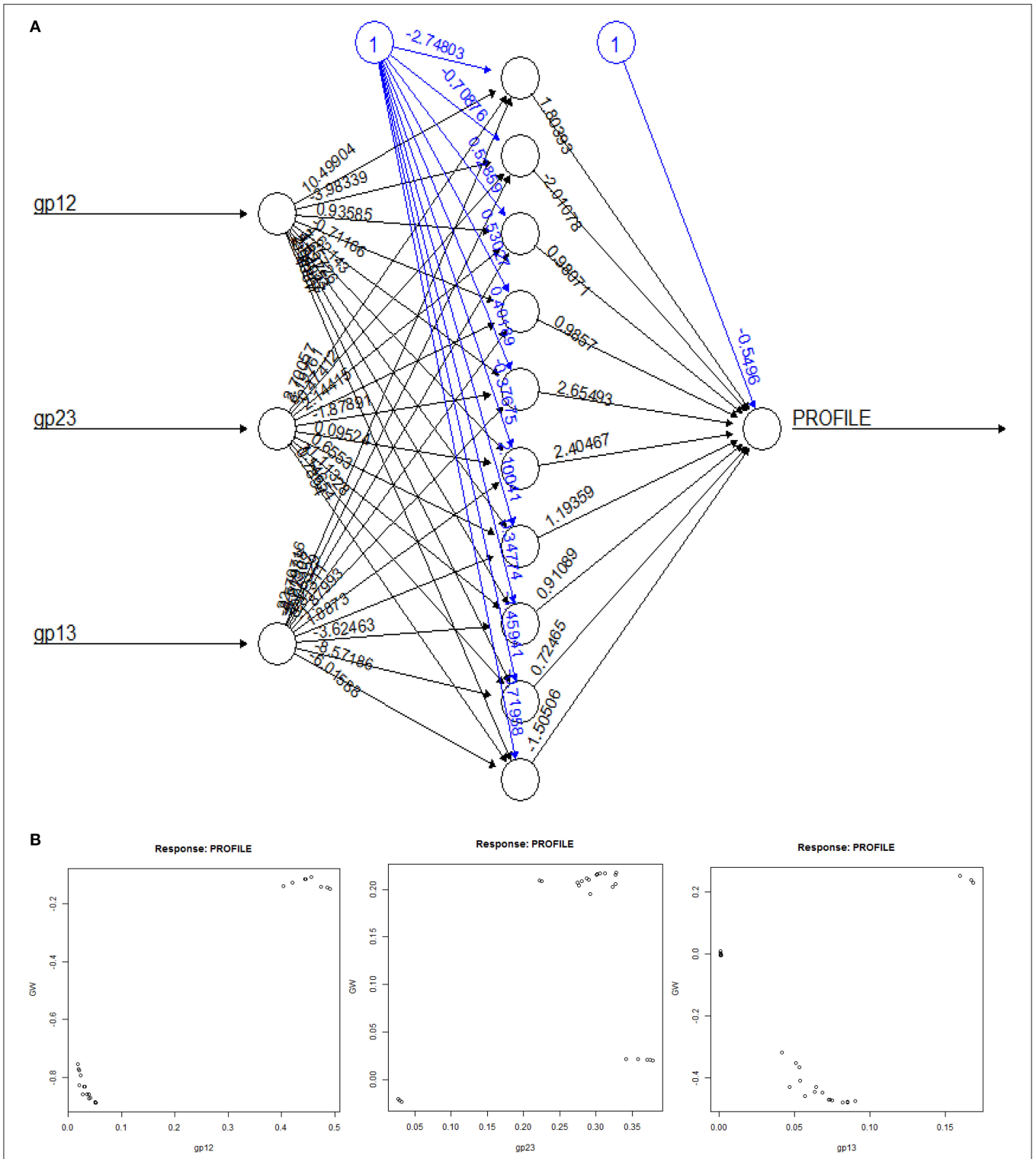



**FIGURE 5 | Details of the predictor-ANN. (A)** The ANN-predicted subclass assignment is dependent on the computed weights and the intercept function. The feedback mechanism for this network is back propagation. **(B)** Plots of $\alpha_{ij}$-scores (gp12, gp23, gp13) the associated weights.

**TABLE 4 | Bounds and conditions for class assignment of GH9 endoglucanases*.**

| Enzyme | Rules |
|---|---|
| Class A | $((9.95 < \beta''_{GH9A} < 10.779) \bigwedge (\alpha_{13} < 0.02))$ |
| Class B | High $3.45 < \beta''_{GH9B} < 5.55$<br>Low $5.55 \leq \beta''_{GH9B} < 8.2052$ |
| Class C | $\left(8.2052 \leq \beta''_{GH9C} \leq 9.95\right) \vee \left((9.95 < \beta''_{GH9C} < 10.779\right) \wedge (\alpha_{13} > 0.02)) \vee (10.779 \leq \beta''_{GH9C} \leq 11.371)$ |

*Dataset used for this definition was $GH9X_{2AA}$.
$\beta''_{GH9A}$, $\beta''_{GH9B}$, $\beta''_{GH9C}$: = ANN-predicted values for filtered sequences of ($GH9X_{2AA} = 92$).

enzymes with predicted class C activity express poorly or not at all (**Figure 7A**). Exceptionally, LOC_Os04g57860 has very high expression in the radicle, with minimal expression in other tissues studied. LOC_Os09g23084 and LOC_Os02g50490, in contrast have a broad expression pattern, with maximum levels of LOC_Os09g23084 mRNA levels observed in the internode pith parenchyma, whole internode, and stigma. The mRNA distribution of the class B endoglucanases ($GH9A_{2AA\_atha} = 2, GH9B_{2AA\_atha} = 8, GH9C_{2AA\_atha} = 4$) is evenly spread with maximum expression in the developing anthers and shoot apical meristem (**Figure 7A**). LOC_Os06g14540, exhibits the highest expression in the plumule, shoot apical meristem, spikelet, palea/lemma, and the developing anthers. LOC_Os04g36610 and LOC_Os09g36060 exhibited very low expression minimal mRNA levels in all the tissues analyzed.

The expression of class B and C genes was also studied in various developmental stages (**Figure 7A**). The transcripts of two class C putative genes LOC_Os09g23084 and LOC_Os02g50490, were mainly detected in the leaf (stages −1, −3), panicle (stages −5, −6), and seed developmental stages (stages 1–4). The remaining class C genes do not seem to express at the levels detected in the developmental stages analyzed. Most of the class B enzymes, except (LOC_Os04g36610, LOC_Os02g50040, and LOC_Os09g36060) on the contrary, exhibit developmental stage-specific expression pattern. The transcripts of LOC_Os06g14540, LOC_Os01g21070, LOC_Os02g50040, and LOC_Os08g02220 accumulate in the early stages of panicle development, whereas, LOC_Os06g14540, LOC_Os01g21070 were mainly detected in the callus suspension, LOC_Os08g29770 and LOC_Os06g14540 mainly express during pre-germination, and the germinating seed stages, respectively. LOC_Os08g29770, has reasonably high mRNA levels during leaf development (stages 1–3), tiller initiation, tillering, and the seed stages (stages 2, 3; **Figure 7A**). The data for *A. thaliana* (**Figure 7B**) suggests that predicted class C genes have similar levels of expression in the anatomical tissues examined, with the loci AT2G32990 and AT1G64390 exhibiting high mRNA levels in the callus, seedling, inflorescence, cell culture, shoot, and root. Class B genes, in contrast, have a poor expression pattern, with the only exceptions being AT4G02290. The stages of maximal gene expression coincide with the stages of callus (AT4G02290 and AT1G71380), inflorescence (AT4G39010, AT4G09740, AT4G39000, and AT4G02290), and root tissues (AT4G02290 andAT1G22880; **Figure 7B**).The expression patterns also suggest that, AT2G32990 and AT1G64390 (class C); AT4G39010 and AT4G02290 (class B), seem to play an important role at all stages of *A. thaliana* development. The class B locus, AT4G39000 exhibits high expression during senescence (**Figure 7B**).

# DISCUSSION

## Unambiguous Assertion of Classes A, B, and C in GH9 Endoglucanses

We have utilized empirical data to identify novel and uncharacterized GH9 endoglucanases from Viridiplantae. The utility of substrates and/or reaction chemistry, structural data, and transcript information to cluster enzymes has been attempted in earlier work, *albeit*, in different biological systems (Kundu, 2012). Since, biochemical information for these enzymes is sparse, we combined available data with well-grounded analytic methods (**Figure 2**) to predict dominant function in GH9 endoglucanases. A general binary classification schema, using a round robin format will convert n-loci into $C_2^n$ pairs, score each, and poll the votes to achieve an overall dominant class (Savicky and Furnkranz, 2003). Since, the unambiguous identification of class C enzymes, mandates, the sequestration of their raw HMM scores, the variance between data pairs, was computed. The ANN prediction was based on the pattern of computed weights for $\alpha_{ij} (= gp_{ij})$ and its equivalence with the cluster mean (Equations 3-5) for each sequence of the training set. Equation 2, may be written as:

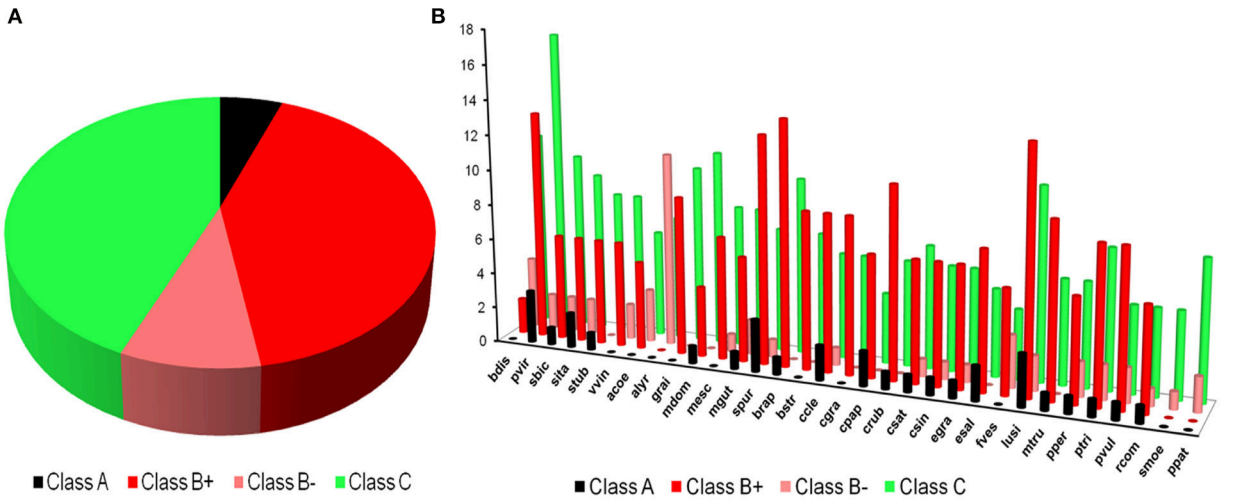
$$\alpha_{ij} = gp_{ij} = \left(|\mu_i - \mu_j|\right)^2 / (100) \left(\sqrt[2]{\left(\sigma_i^2 + \sigma_j^2\right)/\tau}\right) \quad (6)$$

$$\alpha_{ij} \propto 1/\sqrt{\overline{\sigma_{ij}^2}} \quad (7)$$

Clearly, the proportionality constant ($\gamma_{ij}$) for Equation 7, can function as a multiplier ($1 < \gamma_{ij} < \infty$) or as a divisor ($0 < \gamma_{ij} < 1$). This modification, compensates for the inverse relation between the $\alpha_{ij}$-values and the average variance of the relevant data points. Further:

$$\because \beta = \sum \alpha_{ij} = \sum \gamma_{ij} / \sqrt{\overline{\sigma}_{ij}^2} \quad (8)$$

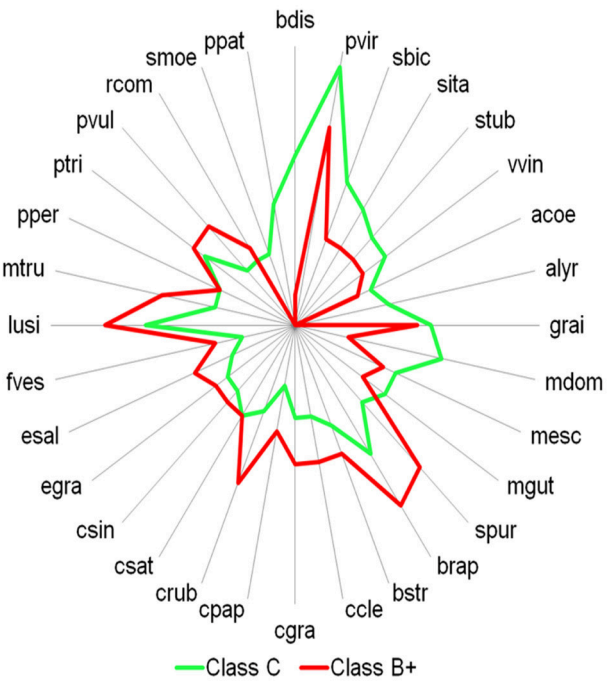
It follows, then, that as the difference between the raw HMM profile scores ($\triangle HMM$) increases, the corresponding $\beta$-value is incremented. The ambiguity in the data trend observed for $GH9X_{2A}$ (**Table 3**) can be interpreted in terms of the inter profile HMM differences ($\triangle HMM$). Thus, for the set ($GH9X_{2AC}$; $\beta < 1.00$), ∼67% of the data was sequestered in the interval $[0, 50)$ or $(0.67)(\{(median\,(\triangle HMM_{GH9AB}) \wedge median\,(\triangle HMM_{GH9BC}) \wedge median\,(\triangle HMM_{GH9AC})\}) < 50$. Similarly, for $\beta > 2.777$, ≈ 86% of the data in $GH9X_{2AA}$belonged to the interval $[150, \infty)$ or $(0.86)(\{(median\,(\triangle HMM_{GH9AB}) \wedge median\,(\triangle HMM_{GH9BC}) \wedge median\,(\triangle HMM_{GH9AC})\}) \geq 150$. However, for intermediate values of $\beta(1.00 \leq \beta < 2.78$ this data, for the highest scoring subset $GH9X_{2AB}$, is heterogeneous,




C

| Taxonomy | | Key | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|---|---|
| VASCULAR PLANTS (N = 31) | MONOCOTS (N = 4) | bdis | | | | | | |
| | | pvir | | | | | | |
| | | sbic | | | | | | |
| | | sita | | | | | | |
| | EUDICOTS (N = 26) | stub | | | | | | |
| | | vvin | | | | | | |
| | | acoe | | | | | | |
| | | alyr | | | | | | |
| | | grai | | | | | | |
| | | mdom | | | | | | |
| | | mesc | | | | | | |
| | | mgut | | | | | | |
| | | spur | | | | | | |
| | | brap | | | | | | |
| | | bstr | | | | | | |
| | | ccle | | | | | | |
| | | cgra | | | | | | |
| | | cpap | | | | | | |
| | | crub | | | | | | |
| | | csat | | | | | | |
| | | csin | | | | | | |
| | | egra | | | | | | |
| | | esal | | | | | | |
| | | fves | | | | | | |
| | | lusi | | | | | | |
| | | mtru | | | | | | |
| | | pper | | | | | | |
| | | ptri | | | | | | |
| | | pvul | | | | | | |
| | | rcom | | | | | | |
| | CLUB-MOSS (N = 1) | smoe | | | | | | |
| MOSSES (N = 1) | | ppat | | | | | | |

D




| | |
|---|---|
| | $GH9A_{2BA}=0$ |
| | $GH9A_{2BA}>0$ |
| | $GH9C_{2BA} \cong GH9B_{2BA}+$ |
| | $GH9C_{2BA} > GH9B_{2BA}+$ |
| | $GH9C_{2BA} < GH9B_{2BA}+$ |
| | $GH9A_{2BA}=0; GH9C_{2BA} < GH9B_{2BA}+$ |
| | $GH9A_{2BA}=0; GH9C_{2BA} \geq GH9B_{2BA}+$ |

**FIGURE 6 | Prediction profile of GH9 endoglucanases in test dataset ($GH9X_{2BA} = 552$). (A)** Generic distribution of sub-class members in the primary proteomes of 32 green plants with sequenced genomes. Here, class B+ refers to sequences with high-confidence in their assignment, in contrast with class B− data. **(B)** Taxonomic spread of classes A, B, and C in selected phytozome data, **(C)** Association table of class specific GH9 proteins, and **(D)** Radar plot of the distribution of members of classes-B+ and C.

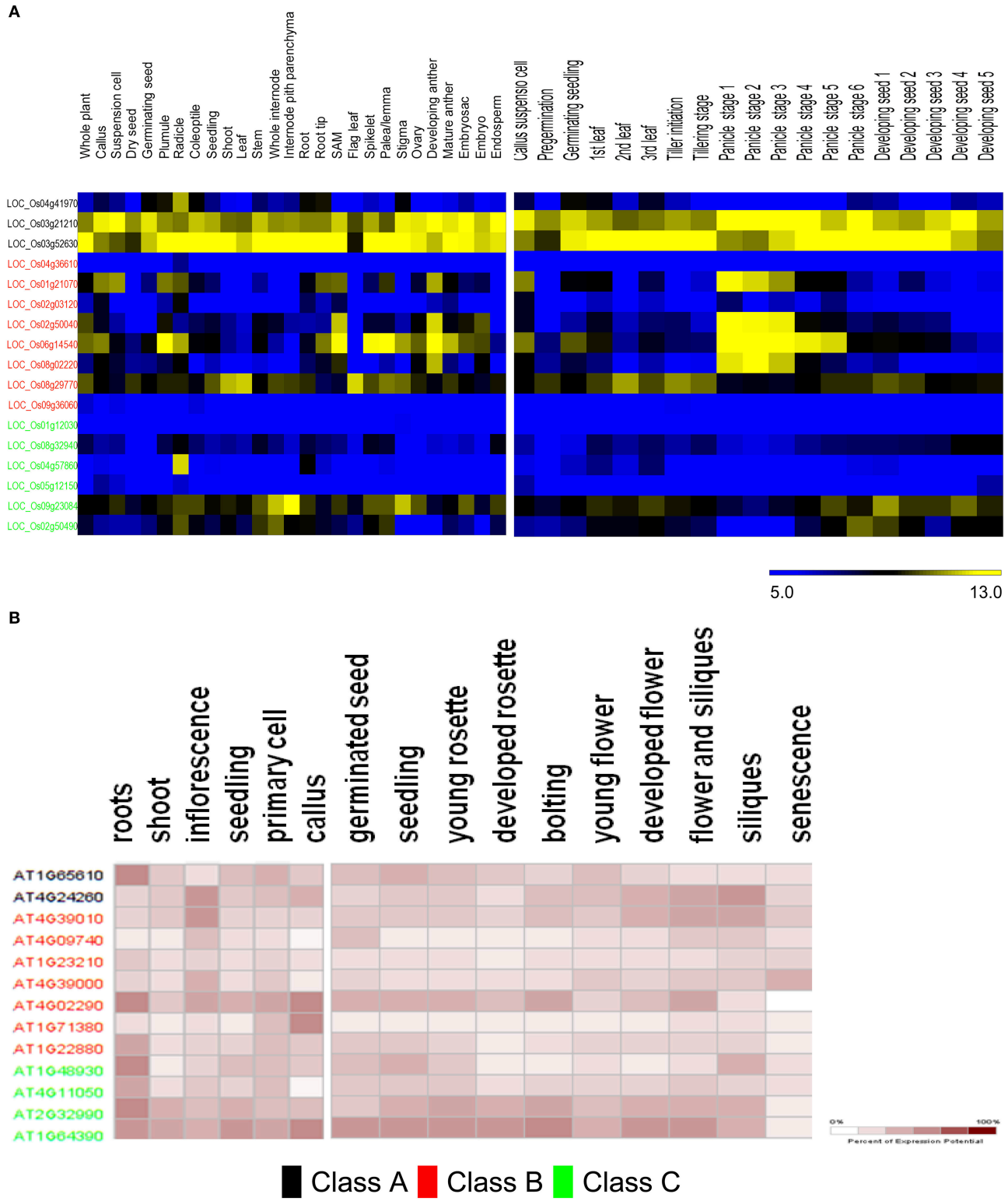



**FIGURE 7 | Meta-analysis of expression data of ANN-predicted GH9 endoglucanses.** Heat maps showing expression patterns of classes A, B, and C in *O. sativa* **(A)** and *A. thaliana* **(B)** in anatomical tissues and at various stages of development. The stages/tissues have been marked at the top. The color bar for *O. sativa* **(A)** represents log2 expression values, blue indicating low-level expression, black medium, and yellow high-level expression, while the same for *A. thaliana* **(B)**, is a gradient from white to maroon and represents percentage of expression potential at the right corner. The downloaded high-resolution images were adapted and presented as such.

**TABLE 5 | Sub-class prediction and distribution of GH9 endoglucanases in Viridiplantae.**

| S. No. | Organism (key) | GH9 members[$] | Sub-class* | ($\beta > 2.777$) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | $GH9A_{2BA}$ | $GH9B_{2BA}$ | | $GH9C_{2BA}$ | Total |
| | | | | | High | Low | | |
| 1. | *Aquilegia coerulea* (acoe) | 23 | 22 | 0 | 5 | 3 | 6 | 14 |
| 2. | *Arabidopsis lyrata* (alyr) | 25 | 25 | 0 | 0 | 11 | 7 | 18 |
| 3. | *Arabidopsis thaliana*[#] (atha) | 26 | 26 | 2 | 8 | 0 | 4 | 14 |
| 4. | *Brassica rapa* FPsc (brap) | 39 | 39 | 1 | 14 | 0 | 10 | 25 |
| 5. | *Brachypodium distachyon* (bdis) | 23 | 23 | 0 | 2 | 4 | 11 | 17 |
| 6. | *Boechera stricta* (bstr) | 23 | 23 | 0 | 9 | 0 | 7 | 16 |
| 7. | *Citrus clementina* (ccle) | 24 | 24 | 2 | 9 | 1 | 6 | 18 |
| 8. | *Capsella grandiflora* (cgra) | 23 | 23 | 0 | 9 | 0 | 6 | 15 |
| 9. | *Carica papaya* (cpap) | 20 | 20 | 2 | 7 | 0 | 4 | 13 |
| 10. | *Capsella rubella* (crub) | 24 | 24 | 1 | 11 | 0 | 6 | 18 |
| 11. | *Cucumis sativus* (csat) | 19 | 19 | 1 | 7 | 1 | 7 | 16 |
| 12. | *Citrus sinensis* (csin) | 23 | 23 | 1 | 7 | 1 | 6 | 15 |
| 13. | *Eucalyptus grandis* (egra) | 29 | 29 | 1 | 7 | 1 | 6 | 15 |
| 14. | *Eutrema salsugineum* (esal) | 24 | 22 | 2 | 8 | 0 | 5 | 22 |
| 15. | *Fragaria vesca* x h (fves) | 23 | 22 | 0 | 6 | 3 | 4 | 13 |
| 16. | *Glycine max*[#] (gmax) | 40 | 39 | 5 | 17 | 0 | 8 | 30 |
| 17. | *Gossypium raimondii* (grai) | 26 | 26 | 0 | 9 | 0 | 10 | 19 |
| 18. | *Linum usitatissimum* (lusi) | 45 | 45 | 3 | 14 | 2 | 11 | 30 |
| 19. | *Malus domestica* (mdom) | 53 | 53 | 1 | 4 | 0 | 11 | 16 |
| 20. | *Manihot esculenta* (mesc) | 28 | 28 | 0 | 7 | 1 | 8 | 16 |
| 21. | *Mimulus guttatus* (mgut) | 23 | 23 | 1 | 6 | 2 | 8 | 17 |
| 22. | *Medicago truncatula* (mtru) | 32 | 32 | 1 | 10 | 0 | 6 | 17 |
| 23. | *Oryza sativa*[#] (osat) | 26 | 26 | 3 | 8 | 0 | 7 | 18 |
| 24. | *Physcomitrella patens* (ppat) | 21 | 21 | 0 | 0 | 2 | 8 | 10 |
| 25. | *Prunus persica* (pper) | 19 | 19 | 1 | 6 | 2 | 6 | 15 |
| 26. | *Populus trichocarpa* (ptri) | 32 | 32 | 1 | 9 | 2 | 8 | 20 |
| 27. | *Panicum virgatum* (pvir) | 50 | 50 | 3 | 13 | 2 | 17 | 35 |
| 28. | *Phaseolus vulgaris* (pvul) | 23 | 23 | 1 | 9 | 2 | 5 | 17 |
| 29. | *Ricinus communis* (rcom) | 21 | 21 | 1 | 6 | 1 | 5 | 13 |
| 30. | *Sorghum bicolor* (sbic) | 25 | 25 | 1 | 6 | 2 | 10 | 19 |
| 31. | *Setaria italica* (sita) | 25 | 25 | 2 | 6 | 2 | 9 | 19 |
| 32. | *Selaginella moellendorfii* (smoe) | 14 | 14 | 0 | 0 | 1 | 5 | 6 |
| 33. | *Salix purpurea* (spur) | 34 | 34 | 3 | 13 | 1 | 7 | 24 |
| 34. | *Solanum lycopersicum*[#] (slyc) | 22 | 22 | 2 | 2 | 0 | 4 | 8 |
| 35. | *Solanum tuberosum* (stub) | 24 | 24 | 1 | 6 | 0 | 8 | 15 |
| 36. | *Vitis vinifera* (vvin) | 23 | 22 | 0 | 6 | 2 | 8 | 16 |
| 37. | *Zea mays*[#] (zmay) | 38 | 38 | 7 | 8 | 0 | 7 | 22 |

[#]*Reference sequences* ($N_2$).

[$]*Selected Phytozome v10.3 sequences were searched using the E5xyD subset of HMMs* ($E_{GH9X} = 10$).

**GH9-homologs were then searched using the E4xyD subset of HMMs* ($E_{GH9A} = E_{GH9B} = E_{GH9C} = 10E\text{-}06$). *The class assignment corresponded to the highest average sequence score with the profile-HMMs.*

*xh, Hybrid.*

*High, Low: Confidence in class B assignment (**Table 2**).*

ambiguous, and spread uniformly ($\approx$ 30–40%) across all the intervals examined (**Table 3** and Table S3C). This data further vindicated our choice of the threshold value.

A comparison with previous predictions of cellulase activity suggests interesting differences. Whilst, sequences with purported class A activity, coincided with earlier work, our annotation attributes dominant class C catalytic activity to a majority of the remainder ($GH9C_{2BA} > GH9B^{+}_{2BA}$) (**Table 5**). This surprising finding, is in complete contrast to earlier predictions, wherein class B enzymes predominate, i.e., $GH9B \gg GH9C$ (Montanier et al., 2010; Buchanan et al., 2012; Xie et al., 2013). Since, the selection of these sequences is

threshold driven, the inclusion of sequences with low confidence scores ($GH9B^{-}_{2BA} = 49$) (**Table 5** and **Figure 6A**) was taken into consideration for some of these calculations. However, despite this, i.e., the total class B enzymes are marginally higher than class C members ($GH9B^{+}_{2BA} + GH9B^{-}_{2BA} > GH9C_{2BA}$; 50% vs.44%). In earlier studies, the overwhelming bias ($GH9B \cong (5)\ (GH9C)$), may be attributed to the indices chosen (sequence similarity and their modifications) to cluster *A. thaliana* and *O. sativa* data (Montanier et al., 2010; Xie et al., 2013). Additionally, since later work on other organisms (*Populus* spp., *Hordeum vulgare*, *Z. Mays*, *Sorghum bicolor*, *Brachypodium distachyon*) used this data as a classification schema, the results were similar (Buchanan et al., 2012; Xie et al., 2013). In our analysis, considerable emphasis has been given to the correlation between the function and organization of class specific sequence or 3D modeled data such as the presence or absence, mutagenesis, and biochemistry of specific regions (secretory peptide, transmembrane, CBM49; **Figure 3C**) in characterized proteins. The resultant class-specific pHMMs with stringent inclusion thresholds, filters, and clustering algorithms, are thus, able to generate noise-free data (distinguish higher- and lower-scoring regions of a particular sequence); thereby generating non conflicting predictions of class A, B, and C enzyme activity (Table S5).

## Function of GH9 Endoglucanases May Vary with Cellulose Content

The role of GH9 endoglucanses in green algae is not clear, and may range from facilitating intercellular/cell matrix adhesion (hydrolysis of cell wall cellulose when present), to extracellular digestion and assimilation of nutrients. Whilst, the presence of cellulose, as a component of the extracellular matrix/cell wall in *Chlorella* (order Trebouxiophyceae), *Oedogonium* (order Chlorophyceae), *Bryopsis* spp. (order Ulvophyceae;), and *C. corallina* (order Charophyceae;) suggest the former; the latter may prevail even in the absence of cell wall cellulose (*C. reinhardtii*, order Chlorophyceae; *O. lucimarinus,* and *M. pussila* spp., order Prasinophyceae; Becker et al., 1989; Estevez et al., 2008; Domozych et al., 2009, 2012; Rodrigues and da Silva Bon, 2011; Blifernez-Klassen et al., 2012; Ciancia et al., 2012). Our results, too, using the generic-HMM ($HMM_{GH9X}$) ignore putative GH9 endoglucanases, from *O. lucimarinus* and *M. pussila* spp. We also observed that green algae spp., descending from *Chlamydomonas* spp. or *Volvox* spp. (*C. reinhardtii, C. subellipsoidea, V. carteri*), despite being selected, by the generic and class specific pHMMs ($HMM_{GH9X}, HMM_{GH9A}, HMM_{GH9B}, HMM_{GH9C}$), had $\beta$-values $< 1.00$. The proximity of HMM scores ($median\ (\triangle HMM) < 25$) (Table S5), too, suggests an overlapping functionality with conflicting biological relevance, thereby, justifying their exclusion.

## Taxonomic Distribution and Expression Pattern of GH9 Endoglucanase May Dictate Dominant Function

The predicted TM domain present in Class A GH9 endoglucanases localizes these enzymes to the membranous compartments of the cell. This suggests that the cellulase activity of this sub-class may be critical to cellulose assembly. In particular, the contribution of this subclass to the formation of a cellulosome, as a protein-carbohydrate connector is, well characterized (Mansoori et al., 2014). Yet, another complementary role for these enzymes is the utilization of the oligosaccharide generated, as a primer. These critical processes in the cellulose based-tiling of the cell wall, clearly are dependent on the focal presence of these endoglucanases. In a related study, class A enzymes were participants in cytokinesis, as well (Zuo et al., 2000). The absence, therefore, of this subclass, in some genera may be expected to retard the biochemical machinery involved in the breakdown of the primary cell wall. Development of uninterrupted xylem cells (absence of cytokinesis), too, could facilitate this. We noticed that a complete absence of class A enzymes ($\approx$ 47.6%), also, interestingly, coincided with increased numbers of class C GH9 endoglucanases in a large number of green plants ($\approx$ 70%; **Figure 6C**). This includes some of the grasses, and other herbaceous plants.

In graminaceae, class C sequences clearly predominate or are at most approximately equal $\left(GH9C_{2BA} \geq GH9B^{+}_{2BA}\right)$ (**Figures 6C,D**). These sequences, at least, hypothetically appear to be more efficient (possess a broader substrate range) as catalysts, a factor which may impede the development of a secondary cell wall, as well as render the primary structure pliable and responsive to abiotic stressors (Kundu, 2015a). Other herbaceous plants such as *V. vinifera, P. patens, S. moellendorffii*, too, possess a non-woody stem, perhaps, secondary to heightened cellulase class C activity. The whole internode, internode pith parenchyma, and developing seed are rich in cellulose content, a factor that may highlight the observations of high mRNA levels in the precursor stem (*O. sativa*; LOC_Os09g23084, LOC_Os02g50490) or shoot (*A. thaliana*; AT2G32990, AT1G64390) regions (**Figure 7**). The abundance of cellulose in some of these tissues suggests, that despite the improved substrate range (crystalline, amorphous), zero order kinetics may predominate in these. This would imply, that putative class C enzymes possess a high Km value, which would render them ineffective when the cellulose content of tissues is minimal (flower, senescence). The reduced cellulose content of the inflorescence and senescent stages (*A. thaliana*) or developing leaf and panicle (*O. sativa*), may require the activities of a biochemically more efficient enzyme (lower Km) with the consequent first order kinetics. Class B enzymes may fulfill this role *in vivo*. Elevated mRNA expression levels of class B genes during these stages in both, of *O. sativa* and *A. thaliana* could support this notion (**Figures 7A,B**).

## Molecular Evolution of Classes B and C Enzymes May Reflect the Development of Complex Physiology

Our findings suggest that the number of GH9 endoglucanases of class B varies inversely with class C enzymes. The non-woody stem of simpler plants suggests a less complex genome organization, thereby, signifying a reduced proteome

with reduced differentiation. As green plants evolved, they incorporated genome segments that coded for proteins of greater complexity and broader functions, the need for an efficient cellulase waned. Thus, class B enzymes appear to frequent the woody, longer living, and more specialized plants, allowing fully developed primary and secondary cell wall structures. The class C identifier, is the CBM49, a 100-120 amino acid region rich in the aromatic and bulky Tryptophan/Tyrosine/Phenylalanine residues. This module is present at the C-terminal end, and is linked by a short stretch of amino acids to the remainder of the protein, which is, in fact, a *de facto* class B sequence (**Figure 3C**; Urbanowicz et al., 2007b). It is possible that, with evolution, this region was spliced out during transcript processing, resulting in the reduced contribution of class C enzymes to general plant physiology, with a reciprocal, dominant presence of class B sequences. Nevertheless, the broad substrate range might compensate, for the poor distribution and/or catalytic efficiency (high Km), thereby, conferring an evolutionary advantage to plants with a functional set of class C GH9 endoglucanases.

### Scope and Limitations of *In silico* Classification of GH9 Endoglucanases

Whilst, accurate, the strength of the assertion of subclass assignment, is dependent on the availability of empirical data (train and validate the HMMs and the ANN), stringency of the sequence filter, and noise-free data (high Signal-to-noise ratio). Clearly, these restrict the utility of the HMM-ANN predictions as a general purpose annotator. Additionally, there is also an in increased loss of information, in terms of sequence(s) elimination (∼7 vs. ∼37%). The integrated pipeline, is also, unlikely to benefit workers with poly-functional enzymes such as the superfamilies of heme-dependent mono- and 2-oxoglutarate dependent di-oxygenases ($N_{pHMM} > 25$) (Kundu, 2012, 2015b). Thus, the enzymes anthocyanidin synthase (EC 1.14.11.19) and clavaminate synthase (EC 1.14.11.21) catalyze substrate hydroxylation and desaturation in tandem, and could, contain overlapping generic 2OG-dependent, hydroxylation, and desaturase profiles. Nevertheless, here too, the subfamilies of the mono-catalytic proline 3- and 4-hydroylases (EC 1.14.11.28, EC 1.14.11.7; EC 1.14.11.2), pentalenolactone- and phytanoic acid-hydroxylases (EC 1.14.11.36; EC 1.14.11.18, might constitute suitable candidates for automatic functional assignment. Since, the HMM-ANN pipeline is dependent on empirical evidence of known function(s), this method may not be suitable as a general sequence annotator.

## CONCLUDING REMARKS

Cellulose digesting GH9 endoglucanases, potentially, have roles in modifying the anatomy and the physiology of plants. The influence on development and response to stress, mandates the pre-emptive breakdown of this glucan. The emergence of plant biomass as a source of biofuel, too, may benefit from the identification of cellulases with a broader substrate range. Alternately, *in vivo* modification could result in plants with abundant and accessible precursor material, facilitating germination, growth, and development. A role for these versatile enzymes, as part of the microbiome promoting biofilms, too, could influence our comprehension of favorable biota for optimal cultivation conditions. However, the limited biochemical characterization of plant proteins (structure, kinetic, mutagenesis), complexity of genetic modification protocols, susceptibility to biotic and abiotic stressors, and heterogeneous growth even in controlled environments, all exert contributory offsets to smooth implementation of these ideas. Despite these, the use of next-generation sequencing, with its precursor genomic data and putative proteome, has the potential to accelerate *in vitro* characterization of computationally predicted functional modules in hypothetical ORFs of plant genomes. Our work, attempts to bridge this divide by constructing a publically available repository of high quality class C plant GH9 endoglucanase sequences.

## AUTHOR CONTRIBUTIONS

SK outlined and designed the study, designed the algorithm for prediction, manually collated all the sequences and their references, carried out the computational analysis, constructed the models, formulated the filters, wrote all necessary code, and the manuscript. RS outlined the study, and participated in revising the manuscript.

## ACKNOWLEDGMENTS

RS gratefully acknowledges the financial support by the Department of Biotechnology, Government of India through the Ramalingaswami fellowship. We also acknowledge financial assistance through DST (Department of Science and Technology)-Purse Grant Phase-II to JNU.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.01185

## REFERENCES

Abbott, D. W., Hrynuik, S., and Boraston, A. B. (2007). Identification and characterization of a novel periplasmic polygalacturonic acid binding protein from *Yersinia enterolitica*. *J. Mol. Biol.* 367, 1023–1033. doi: 10.1016/j.jmb.2007.01.030

Abramyan, J., and Stajich, J. E. (2012). Species-specific chitin-binding module 18 expansion in the amphibian pathogen *Batrachochytrium dendrobatidis*. *mBio* 3, e00150–e00112. doi: 10.1128/mBio.00150-12

Agarwal, V., Dauenhauer, P. J., Huber, G. W., and Auerbach, S. M. (2012). Ab initio dynamics of cellulose pyrolysis: nascent decomposition pathways at 327 and 600 degrees C. *J. Am. Chem. Soc.* 134, 14958–14972. doi: 10.1021/ja305135u

Alahuhta, M., Xu, Q., Bomble, Y. J., Brunecky, R., Adney, W. S., Ding, S. Y., et al. (2010). The unique binding mode of cellulosomal CBM4 from *Clostridium thermocellum* cellobiohydrolase A. *J. Mol. Biol.* 402, 374–387. doi: 10.1016/j.jmb.2010.07.028

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., et al. (2000). InterPro Consortium. InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16, 1145–1150. doi: 10.1093/bioinformatics/16.12.1145

Artzi, L., Morag, E., Shamshoum, M., and Bayer, E. A. (2016). Cellulosomal expansin: functionality and incorporation into the complex. *Biotechnol. Biofuels* 9, 61. doi: 10.1186/s13068-016-0474-5

Augimeri, R. V., Varley, A. J., and Strap, J. L. (2015). Establishing a role for bacterial cellulose in environmental interactions: lessons learned from diverse biofilm-producing *Proteobacteria*. *Front. Microbiol.* 6:1282. doi: 10.3389/fmicb.2015.01282

Bachman, E. S., and McClay, D. R. (1996). Molecular cloning of the first metazoan beta-1,3 glucanase from eggs of the sea urchin *Strongylocentrotus purpuratus*. *Proc. Natl. Acad. Sci. U.S.A.* 93, 6808–6813. doi: 10.1073/pnas.93.13.6808

Barral, P., Suárez, C., Batanero, E., Alfonso, C., Alché Jde, D., Rodríguez-García, M. I., et al. (2005). An olive pollen protein with allergenic activity, Ole e 10, defines a novel family of carbohydrate-binding modules and is potentially implicated in pollen germination. *Biochem. J.* 390, 77–84. doi: 10.1042/BJ20050456

Becker, B., Hård, K., Melkonian, M., Kamerling, J. P., and Vliegenthart, J. F. (1989). Identification of 3-deoxy-manno-2-octulosonic acid, 3-deoxy-5-O-methyl-manno-2-octulosonic acid and 3-deoxy-lyxo-2-heptulosaric acid in the cell wall (theca) of the green alga Tetraselmis striata Butcher (*Prasinophyceae*). *Eur. J. Biochem.* 182, 153–160. doi: 10.1111/j.1432-1033.1989.tb14811.x

Blifernez-Klassen, O., Klassen, V., Doebbe, A., Kersting, K., Grimm, P., Wobbe, L., et al. (2012). Cellulose degradation and assimilation by the unicellular phototrophic eukaryote *Chlamydomonas reinhardtii*. *Nat. Commun.* 3, 1214. doi: 10.1038/ncomms2210

Blume, J. E., and Ennis, H. L. (1991). A *Dictyostelium discoideum* cellulase is a member of a spore germination-specific gene family. *J. Biol. Chem.* 266, 15432–15437.

Boraston, A. B., Bolam, D. N., Gilbert, H. J., and Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* 382, 769–781. doi: 10.1042/BJ20040892

Boraston, A. B., Kwan, E., Chiu, P., Warren, R. A., and Kilburn, D. G. (2003). Recognition and hydrolysis of noncrystalline cellulose. *J. Biol. Chem.* 278, 6120–6127. doi: 10.1074/jbc.M209554200

Brummell, D. A., Catala, C., Lashbrook, C. C., and Bennett, A. B. (1997). A membrane-anchored E-type endo-1,4-beta-glucanase is localized on Golgi and plasma membranes of higher plants. *Proc. Natl. Acad. Sci. U.S.A.* 94, 4794–4799. doi: 10.1073/pnas.94.9.4794

Buchanan, M., Burton, R. A., Dhugga, K. S., Rafalski, A. J., Tingey, S. V., Shirley, N. J., et al. (2012). Endo-(1,4)-beta-glucanase gene families in the grasses: temporal and spatial co-transcription of orthologous genes. *BMC Plant Biol.* 12:235. doi: 10.1186/1471-2229-12-235

Cannon, R. E., and Anderson, S. M. (1991). Biogenesis of bacterial cellulose. *Crit. Rev. Microbiol.* 17, 435–447. doi: 10.3109/10408419109115207

Cao, P., Jung, K. H., Choi, D., Hwang, D., Zhu, J., and Ronald, P. C. (2012). The Rice Oligonucleotide Array Database: an atlas of rice gene expression. *Rice* 5:17. doi: 10.1186/1939-8433-5-17

Catala, C., and Bennett, A. B. (1998). Cloning and sequence analysis of TomCel8; a new plantendo-beta-1,4-D-glucanase gene, encoding a protein with a putative carbohydrate binding domain (Accession No. AF098292) (PGR98-209). *Plant Physiol.* 118, 1535.

Catalá, C., Rose, J. K., and Bennett, A. B. (1997). Auxin regulation and spatial localization of an endo-1,4-beta-D-glucanase and a xyloglucan endotransglycosylase in expanding tomato hypocotyls. *Plant J.* 12, 417–426. doi: 10.1046/j.1365-313X.1997.12020417.x

Chung, D., Young, J., Cha, M., Brunecky, R., Bomble, Y. J., Himmel, M. E., et al. (2015). Expression of the Acidothermus cellulolyticus E1 endoglucanase in Caldicellulosiruptor bescii enhances its ability to deconstruct crystalline cellulose. *Biotechnol. Biofuels* 8, 113. doi: 10.1186/s13068-015-0296-x

Ciancia, M., Alberghina, J., Arata, P. X., Benavides, H., Leliaert, F., Verbruggen, H., et al. (2012). Characterization of Cell Wall Polysaccharides of the Coencocytic Green Seaweed Bryopsis Plumosa (*Bryopsidaceae*, *Chlorophyta*) from the Argentine Coast(1). *J. Phycol.* 48, 326–335. doi: 10.1111/j.1529-8817.2012.01131.x

del Campillo, E., Gaddam, S., Mettle-Amuah, D., and Heneks, J. (2012). A tale of two tissues: AtGH9C1 is an endo-beta-1,4-glucanase involved in root hair and endosperm development in Arabidopsis. *PLoS ONE* 7:e49363. doi: 10.1371/journal.pone.0049363

Domozych, D. S., Ciancia, M., Fangel, J. U., Mikkelsen, M. D., Ulvskov, P., and Willats, W. G. (2012). The cell walls of green algae: a journey through evolution and diversity. *Front. Plant Sci.* 3:82. doi: 10.3389/fpls.2012.00082

Domozych, D. S., Sørensen, I., and Willats, W. G. (2009). The distribution of cell wall polymers during antheridium development and spermatogenesis in the Charophycean green alga, Chara corallina. *Ann. Bot.* 104, 1045–1056. doi: 10.1093/aob/mcp193

Estevez, J. M., Leonardi, P. I., and Alberghina, J. S. (2008). Cell Wall Carbohydrate Epitopes in the Green Alga Oedogonium Bharuchae, F. Minor (Oedogoniales, *Chlorophyta*)(1). *J. Phycol.* 44, 1257–1268. doi: 10.1111/j.1529-8817.2008.00568.x

Ficko-Blean, E., and Boraston, A. B. (2006). The interaction of a carbohydrate-binding module from a *Clostridium perfringens* N-acetyl-beta-hexosaminidase with its carbohydrate receptor. *J. Biol. Chem.* 281, 37748–37757. doi: 10.1074/jbc.M606126200

Flint, J., Bolam, D. N., Nurizzo, D., Taylor, E. J., Williamson, M. P., Walters, C., et al. (2005). Probing the mechanism of ligand recognition in family 29 carbohydrate-binding modules. *J. Biol. Chem.* 280, 23718–23726. doi: 10.1074/jbc.M501551200

Gille, C., Fähling, M., Weyand, B., Wieland, T., and Gille, A. (2014). Alignment-Annotator web server: rendering and annotating sequence alignments. *Nucleic Acids Res.* 42, W3–W6. doi: 10.1093/nar/gku400

Goellner, M., Wang, X., and Davis, E. L. (2001). Endo-beta-1,4-glucanase expression in compatible plant-nematode interactions. *Plant Cell* 13, 2241–2255. doi: 10.1105/tpc.13.10.2241

Jamal, S., Nurizzo, D., Boraston, A. B., and Davies, G. J. (2004). X-ray crystal structure of a non-crystalline cellulose-specific carbohydrate-binding module: CBM28. *J. Mol. Biol.* 339, 253–258. doi: 10.1016/j.jmb.2004.03.069

Janecek, Š., Svensson, B., and MacGregor, E. A. (2011). Structural and evolutionary aspects of two families of non-catalytic domains present in starch and glycogen binding proteins from microbes, plants and animals. *Enzyme Microb. Technol.* 49, 429–440. doi: 10.1016/j.enzmictec.2011.07.002

Kalaitzis, P., Hong, S. B., Solomos, T., and Tucker, M. L. (1999). Molecular characterization of a tomato endo-beta-1,4-glucanase gene expressed in mature pistils, abscission zones and fruit. *Plant Cell Physiol.* 40, 905–908. doi: 10.1093/oxfordjournals.pcp.a029621

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Khater, S., and Mohanty, D. (2015). In silico identification of AMPylating enzymes and study of their divergent evolution. *Sci. Rep.* 5:10804. doi: 10.1038/srep10804

Klemm, D., Heublein, B., Fink, H. P., and Bohn, A. (2005). Cellulose: fascinating biopolymer and sustainable raw material. *Angew. Chem. Int. Ed. Engl.* 44, 3358–3393. doi: 10.1002/anie.200460587

Koseki, T., Mese, Y., Fushinobu, S., Masaki, K., Fujii, T., Ito, K., et al. (2008). Biochemical characterization of a glycoside hydrolase family 61 endoglucanase from *Aspergillus kawachii*. *Appl. Microbiol. Biotechnol.* 77, 1279–1285. doi: 10.1007/s00253-007-1274-4

Kundu, S. (2012). Distribution and prediction of catalytic domains in 2-oxoglutarate dependent dioxygenases. *BMC Res. Notes* 5:410. doi: 10.1186/1756-0500-5-410

Kundu, S. (2015a). Co-operative intermolecular kinetics of 2-oxoglutarate dependent dioxygenases may be essential for system-level regulation of plant cell physiology. *Front. Plant Sci.* 6:489. doi: 10.3389/fpls.2015.00489

Kundu, S. (2015b). Unity in diversity, a systems approach to regulating plant cell physiology by 2-oxoglutarate-dependent dioxygenases. *Front. Plant Sci.* 6:98. doi: 10.3389/fpls.2015.00098

Lane, D. R., Wiedemeier, A., Peng, L., Höfte, H., Vernhettes, S., Desprez, T., et al. (2001). Temperature-sensitive alleles of RSW2 link the KORRIGAN endo-1,4-beta-glucanase to cellulose synthesis and cytokinesis in Arabidopsis. *Plant Physiol.* 126, 278–288. doi: 10.1104/pp.126.1.278

Lashbrook, C. C., Gonzalez-Bosch, C., and Bennett, A. B. (1994). Two divergent endo-beta-1,4-glucanase genes exhibit overlapping expression in ripening fruit and abscising flowers. *Plant Cell* 6, 1485–1493. doi: 10.1105/tpc.6.10.1485

Lees, J., Yeats, C., Redfern, O., Clegg, A., and Orengo, C. (2010). Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.* 38, D296–D300. doi: 10.1093/nar/gkp987

Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., et al. (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30, 242–244. doi: 10.1093/nar/30.1.242

Li, S., Yang, X., Bao, M., Wu, Y., Yu, W., and Han, F. (2015). Family 13 carbohydrate-binding module of alginate lyase from Agarivorans sp. L11 enhances its catalytic efficiency and thermostability, and alters its substrate preference and product distribution. *FEMS Microbiol. Lett.* 362:fnv054. doi: 10.1093/femsle/fnv054

Libertini, E., Li, Y., and McQueen-Mason, S. J. (2004). Phylogenetic analysis of the plant endo-beta-1,4-glucanase gene family. *J. Mol. Evol.* 58, 506–515. doi: 10.1007/s00239-003-2571-x

Llop-Tous, I., Domínguez-Puigjaner, E., Palomer, X., and Vendrell, M. (1999). Characterization of two divergent endo-beta-1,4-glucanase cDNA clones highly expressed in the nonclimacteric strawberry fruit. *Plant Physiol.* 119, 1415–1422. doi: 10.1104/pp.119.4.1415

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

Lopez-Casado, G., Urbanowicz, B. R., Damasceno, C. M., and Rose, J. K. (2008). Plant glycosyl hydrolases and biofuels: a natural marriage. *Curr. Opin. Plant Biol.* 11, 329–337. doi: 10.1016/j.pbi.2008.02.010

Maloney, V. J., and Mansfield, S. D. (2010). Characterization and varied expression of a membrane-bound endo-beta-1,4-glucanase in hybrid poplar. *Plant Biotechnol. J.* 8, 294–307. doi: 10.1111/j.1467-7652.2009.00483.x

Maloney, V. J., Samuels, A. L., and Mansfield, S. D. (2012). The endo-1,4-beta-glucanase Korrigan exhibits functional conservation between gymnosperms and angiosperms and is required for proper cell wall formation in gymnosperms. *New Phytol.* 193, 1076–1087. doi: 10.1111/j.1469-8137.2011.03998.x

Mansoori, N., Timmers, J., Desprez, T., Alvim-Kamei, C. L., Dees, D. C., Vincken, J. P., et al. (2014). KORRIGAN1 interacts specifically with integral components of the cellulose synthase machinery. *PLoS ONE* 9:e112387. doi: 10.1371/journal.pone.0112387

Mathee, K., Ciofu, O., Sternberg, C., Lindum, P. W., Campbell, J. I., Jensen, P., et al. (1999). Mucoid conversion of *Pseudomonas aeruginosa* by hydrogen peroxide: a mechanism for virulence activation in the cystic fibrosis lung. *Microbiology* 145(Pt 6), 1349–1357. doi: 10.1099/13500872-145-6-1349

Matthysse, A. G., Marry, M., Krall, L., Kaye, M., Ramey, B. E., Fuqua, C., et al. (2005). The effect of cellulose overproduction on binding and biofilm formation on roots by *Agrobacterium tumefaciens*. *Mol. Plant Microbe Interact.* 18, 1002–1010. doi: 10.1094/MPMI-18-1002

Matthysse, A. G., Thomas, D. L., and White, A. R. (1995). Mechanism of cellulose synthesis in *Agrobacterium tumefaciens*. *J. Bacteriol.* 177, 1076–1081.

McLean, B. W., Bray, M. R., Boraston, A. B., Gilkes, N. R., Haynes, C. A., and Kilburn, D. G. (2000). Analysis of binding of the family 2a carbohydrate-binding module from *Cellulomonas fimi* xylanase 10A to cellulose: specificity and identification of functionally important amino acid residues. *Protein Eng.* 13, 801–809. doi: 10.1093/protein/13.11.801

Milligan, S. B., and Gasser, C. S. (1995). Nature and regulation of pistil-expressed genes in tomato. *Plant Mol. Biol.* 28, 691–711. doi: 10.1007/BF00021194

Mølhøj, M., Pagant, S., and Höfte, H. (2002). Towards understanding the role of membrane-bound endo-beta-1,4-glucanases in cellulose biosynthesis. *Plant Cell Physiol.* 43, 1399–1406. doi: 10.1093/pcp/pcf163

Montanier, C., Flint, J. E., Bolam, D. N., Xie, H., Liu, Z., Rogowski, A., et al. (2010). Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. *J. Biol. Chem.* 285, 31742–31754. doi: 10.1074/jbc.M110.142133

Nakamura, S., Mori, H., Sakai, F., and Hayashi, T. (1995). Cloning and sequencing of a cDNA for poplar endo-1,4-beta-glucanase. *Plant Cell Physiol.* 36, 1229–1235.

Newstead, S. L., Watson, J. N., Bennet, A. J., and Taylor, G. (2005). Galactose recognition by the carbohydrate-binding module of a bacterial sialidase. *Acta Crystallogr. D Biol. Crystallogr.* 61, 1483–1491. doi: 10.1107/S0907444905026132

Nicol, F., His, I., Jauneau, A., Vernhettes, S., Canut, H., and Höfte, H. (1998). A plasma membrane-bound putative endo-1,4-beta-D-glucanase is required for normal wall assembly and cell elongation in Arabidopsis. *EMBO J.* 17, 5563–5576. doi: 10.1093/emboj/17.19.5563

Obembe, O. O., Jacobsen, E., Timmers, J., Gilbert, H., Blake, A. W., Knox, J. P., et al. (2007). Promiscuous, non-catalytic, tandem carbohydrate-binding modules modulate the cell-wall structure and development of transgenic tobacco (*Nicotiana tabacum*) plants. *J. Plant Res.* 120, 605–617. doi: 10.1007/s10265-007-0099-7

Otegui, M. S. (2007). "Endosperm cell walls: formation, composition, and functions," in *Endosperm: Development and Molecular Biology*, ed O.-A. Olsen (Heidelberg: Springer-Verlag), 159–177.

Palomo, M., Kralj, S., van der Maarel, M. J., and Dijkhuizen, L. (2009). The unique branching patterns of Deinococcus glycogen branching enzymes are determined by their N-terminal domains. *Appl. Environ. Microbiol.* 75, 1355–1362. doi: 10.1128/AEM.02141-08

Paulsen, A. D., Hough, B. R., Williams, C. L., Teixeira, A. R., Schwartz, D. T., Pfaendtner, J., et al. (2014). Fast pyrolysis of wood for biofuels: spatiotemporally resolved diffuse reflectance in situ spectroscopy of particles. *ChemSusChem* 7, 765–776. doi: 10.1002/cssc.201301056

Peng, L., Kawagoe, Y., Hogan, P., and Delmer, D. (2002). Sitosterol-beta-glucoside as primer for cellulose synthesis in plants. *Science* 295, 147–150. doi: 10.1126/science.1064281

Reardon-Robinson, M. E., Wu, C., Mishra, A., Chang, C., Bier, N., Das, A., et al. (2014). Pilus hijacking by a bacterial coaggregation factor critical for oral biofilm development. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3835–3840. doi: 10.1073/pnas.1321417111

Rensink, W. A., Lee, Y., Liu, J., Iobst, S., Ouyang, S., and Buell, C. R. (2005). Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. *BMC Genomics* 6:124. doi: 10.1186/1471-2164-6-124

Robledo, M., Rivera, L., Jiménez-Zurdo, J. I., Rivas, R., Dazzo, F., Velázquez, E., et al. (2012). Role of Rhizobium endoglucanase CelC2 in cellulose biosynthesis and biofilm formation on plant roots and abiotic surfaces. *Microb. Cell Fact.* 11:125. doi: 10.1186/1475-2859-11-125

Rodrigues, M. A., and da Silva Bon, E. P. (2011). Evaluation of Chlorella (*Chlorophyta*) as source of fermentable sugars via cell wall enzymatic hydrolysis. *Enzyme Res.* 2011:405603. doi: 10.4061/2011/405603

Roske, Y., Sunna, A., Pfeil, W., and Heinemann, U. (2004). High-resolution crystal structures of Caldicellulosiruptor strain Rt8B.4 carbohydrate-binding module CBM27-1 and its complex with mannohexaose. *J. Mol. Biol.* 340, 543–554. doi: 10.1016/j.jmb.2004.04.072

Sato, S., Kato, T., Kakegawa, K., Ishii, T., Liu, Y. G., Awano, T., et al. (2001). Role of the putative membrane-bound endo-1,4-beta-glucanase KORRIGAN in cell elongation and cellulose synthesis in *Arabidopsis thaliana*. *Plant Cell Physiol.* 42, 251–263. doi: 10.1093/pcp/pce045

Savicky, P., and Furnkranz, J. (2003). Combining pairwise classifiers with stacking. *Adv. Intell. Data Anal. V* 2810, 219–229. doi: 10.1007/978-3-540-45231-7_21

Shani, Z., Dekel, M., Roiz, L., Horowitz, M., Kolosovski, N., Lapidot, S., et al. (2006). Expression of endo-1,4-beta-glucanase (cel1) in *Arabidopsis thaliana* is associated with plant growth, xylem development and cell wall thickening. *Plant Cell Rep.* 25, 1067–1074. doi: 10.1007/s00299-006-0167-9

Shani, Z., Dekel, M., Tsabary, G., and Shoseyov, O. (1997). Cloning and characterization of elongation specific endo-1,4-beta-glucanase (cel1) from *Arabidopsis thaliana*. *Plant Mol. Biol.* 34, 837–842. doi: 10.1023/A:1005849627301

Sharma, R., Cao, P., Jung, K. H., Sharma, M. K., and Ronald, P. C. (2013). Construction of a rice glycoside hydrolase phylogenomic database and identification of targets for biofuel research. *Front. Plant Sci.* 4:330. doi: 10.3389/fpls.2013.00330

Shpigel, E., Roiz, L., Goren, R., and Shoseyov, O. (1998). Bacterial cellulose-binding domain modulates *in vitro* elongation of different plant cells. *Plant Physiol.* 117, 1185–1194. doi: 10.1104/pp.117.4.1185

Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a protein domain database for functional

characterization and annotation. *Nucleic Acids Res.* 38, D161–D166. doi: 10.1093/nar/gkp885

Simpson, P. J., Xie, H., Bolam, D. N., Gilbert, H. J., and Williamson, M. P. (2000). The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J. Biol. Chem.* 275, 41137–41142. doi: 10.1074/jbc.M006948200

Skibbe, D. S., Wang, X., Zhao, X., Borsuk, L. A., Nettleton, D., and Schnable, P. S. (2006). Scanning microarrays at multiple intensities enhances discovery of differentially expressed genes. *Bioinformatics* 22, 1863–1870. doi: 10.1093/bioinformatics/btl270

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26, 320–322. doi: 10.1093/nar/26.1.320

Sturcova, A., His, I., Apperley, D. C., Sugiyama, J., and Jarvis, M. C. (2004). Structural details of crystalline cellulose from higher plants. *Biomacromolecules* 5, 1333–1339. doi: 10.1021/bm034517p

Trainotti, L., Spolaore, S., Ferrarese, L., and Casadoro, G. (1997). Characterization of ppEG1, a member of a multigene family which encodes endo-beta-1,4-glucanase in peach. *Plant Mol. Biol.* 34, 791–802.

Tsabary, G., Shani, Z., Roiz, L., Levy, I., Riov, J., and Shoseyov, O. (2003). Abnormal 'wrinkled' cell walls and retarded development of transgenic Arabidopsis thaliana plants expressing endo-1,4-beta-glucanase (cell) antisense. *Plant Mol. Biol.* 51, 213–224. doi: 10.1023/A:1021162321527

Tucker, M. L., and Milligan, S. B. (1991). Sequence analysis and comparison of avocado fruit and bean abscission cellulases. *Plant Physiol.* 95, 928–933. doi: 10.1104/pp.95.3.928

Tucker, M. L., Sexton, R., Del Campillo, E., and Lewis, L. N. (1988). Bean abscission cellulase: characterization of a cDNA clone and regulation of gene expression by ethylene and auxin. *Plant Physiol.* 88, 1257–1262. doi: 10.1104/pp.88.4.1257

Tunnicliffe, R. B., Bolam, D. N., Pell, G., Gilbert, H. J., and Williamson, M. P. (2005). Structure of a mannan-specific family 35 carbohydrate-binding module: evidence for significant conformational changes upon ligand binding. *J. Mol. Biol.* 347, 287–296. doi: 10.1016/j.jmb.2005.01.038

Uni, F., Lee, S., Yatsunami, R., Fukui, T., and Nakamura, S. (2009). Role of exposed aromatic residues in substrate-binding of CBM family 5 chitin-binding domain of alkaline chitinase. *Nucleic Acids Symp. Ser.* 53, 311–312. doi: 10.1093/nass/nrp156

Updegraff, D. M. (1969). Semimicro determination of cellulose in biological materials. *Anal. Biochem.* 32, 420–424. doi: 10.1016/S0003-2697(69)80009-6

Urbanowicz, B. R., Bennett, A. B., Del Campillo, E., Catalá, C., Hayashi, T., Henrissat, B., et al. (2007a). Structural organization and a standardized nomenclature for plant endo-1,4-beta-glucanases (cellulases) of glycosyl hydrolase family 9. *Plant Physiol.* 144, 1693–1696. doi: 10.1104/pp.107.102574

Urbanowicz, B. R., Catalá, C., Irwin, D., Wilson, D. B., Ripoll, D. R., and Rose, J. K. (2007b). A tomato endo-beta-1,4-glucanase, SlCel9C1, represents a distinct subclass with a new family of carbohydrate binding modules (CBM49). *J. Biol. Chem.* 282, 12066–12074. doi: 10.1074/jbc.M607925200

White, A. P., Gibson, D. L., Collinson, S. K., Banser, P. A., and Kay, W. W. (2003). Extracellular polysaccharides associated with thin aggregative fimbriae of *Salmonella enterica* serovar enteritidis. *J. Bacteriol.* 185, 5398–5407. doi: 10.1128/JB.185.18.5398-5407.2003

Wu, S. C., Blumer, J. M., Darvill, A. G., and Albersheim, P. (1996). Characterization of an endo-beta-1,4-glucanase gene induced by auxin in elongating pea epicotyls. *Plant Physiol.* 110, 163–170. doi: 10.1104/pp.110.1.163

Xie, G., Yang, B., Xu, Z., Li, F., Guo, K., Zhang, M., et al. (2013). Global identification of multiple OsGH9 family members and their involvement in cellulose crystallinity modification in rice. *PLoS ONE* 8:e50171. doi: 10.1371/journal.pone.0050171

Yoshida, K., Imaizumi, N., Kaneko, S., Kawagoe, Y., Tagiri, A., Tanaka, H., et al. (2006). Carbohydrate-binding module of a rice endo-beta-1,4-glycanase, OsCel9A, expressed in auxin-induced lateral root primordia, is post-translationally truncated. *Plant Cell Physiol.* 47, 1555–1571. doi: 10.1093/pcp/pcl021

Yoshida, M., Igarashi, K., Wada, M., Kaneko, S., Suzuki, N., Matsumura, H., et al. (2005). Characterization of carbohydrate-binding cytochrome b562 from the white-rot fungus *Phanerochaete Chrysosporium*. *Appl. Environ. Microbiol.* 71, 4548–4555. doi: 10.1128/AEM.71.8.4548-4555.2005

Yoshida, Y., Palmer, R. J., Yang, J., Kolenbrander, P. E., and Cisar, J. O. (2006). Streptococcal receptor polysaccharides: recognition molecules for oral biofilm formation. *BMC Oral Health*. 6(Suppl. 1):S12. doi: 10.1186/1472-6831-6-S1-S12

Yu, L., Chen, H., Sun, J., and Li, L. (2014). PtrKOR1 is required for secondary cell wall cellulose biosynthesis in Populus. *Tree Physiol.* 34, 1289–1300. doi: 10.1093/treephys/tpu020

Yung, M. H., Schaffer, R., and Putterill, J. (1999). Identification of genes expressed during early Arabidopsis carpel development by mRNA differential display: characterisation of ATCEL2, a novel endo-1,4-beta-D-glucanase gene. *Plant J.* 17, 203–208. doi: 10.1046/j.1365-313X.1999.00359.x

Zhang, Q., Zhang, X., Wang, P., Li, D., Chen, G., Gao, P., et al. (2015). Determination of the action modes of cellulases from hydrolytic profiles over a time course using fluorescence-assisted carbohydrate electrophoresis. *Electrophoresis* 36, 910–917. doi: 10.1002/elps.201400563

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol.* 136, 2621–2632. doi: 10.1104/pp.104.046367

Zogaj, X., Nimtz, M., Rohde, M., Bokranz, W., and Römling, U. (2001). The multicellular morphotypes of Salmonella typhimurium and Escherichia coli produce cellulose as the second component of the extracellular matrix. *Mol. Microbiol.* 39, 1452–1463. doi: 10.1046/j.1365-2958.2001.02337.x

Zuo, J., Niu, Q. W., Nishizawa, N., Wu, Y., Kost, B., and Chua, N. H. (2000). KORRIGAN, an Arabidopsis endo-1,4-beta-glucanase, localizes to the cell plate by polarized targeting and is essential for cytokinesis. *Plant Cell* 12, 1137–1152. doi: 10.1105/tpc.12.7.1137

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.