The Evolutionary History and Dynamics of the Cellulose Synthase Superfamily

Julian George Schwerdt

School of Agriculture, Food and Wine

University of Adelaide

A thesis submitted in fulfillment of the requirements for

the degree of Doctor of Philosophy

August, 2016

Table of Contents

Abstract

The plant cell wall is central to the success of the embryophyte radiation. The high tensile strength of the cell wall supports complex branching architectures adapted to a varying and highly competitive environment. The cell wall has also played an integral role during the evolution of multicellularity by bonding cells together, controlling cell differentiation, acting as an energy store and mediating chemical signals. Polysaccharides are the dominant component of the plant cell wall and the genes involved in their biosynthesis are a major focus of cell wall research. The work presented in this thesis aims to reconstruct the evolutionary history and selection dynamics of the embryophyte cellulose synthase (*CesA*) and cellulose synthase-like (*Csl*) superfamily.

The commercially significant Poaceae (grasses) have received considerable attention. The commercially significant Poaceae (grasses) have received considerable attention from the plant cell wall research community, not least because they are unique in containing a high abundance of (1,3;1,4)-β-glucan. Chapter 2 reconstructs the molecular phylogeny and evolutionary dynamics of the *CesA* superfamily in the Poaceae. Bayesian and likelihood-based models yielded a well-resolved gene tree for the superfamily and revealed heterogeneous selection pressures among amino acid sites. To provide a functional context to these findings, an energetically refined homology model of HvCslF6 was constructed — this is an important enzyme implicated in the biosynthesis of (1,3;1,4)-β-glucan — that was used to map amino-acid residues under selection onto a three-dimensional structure.

Analyses performed for Chapter 2 showed that the *CslJ* clade was conspicuous in having a level of historical divergence too high for the evolutionary models used. As high divergence could indicate functional shift, the focus in Chapter 3 was on the phylogenetic analysis and functional characterisation of *CslJ*. Phylogenetic analyses of *CslE, CslJ* and *CslG* families across an improved taxonomic sampling of fully sequenced eudicot and monocot species were performed and experimental evidence that *CslJ* is implicated in the biosynthesis of (1,3;1,4)-β-glucan is presented. Selection tests show that the *CslJ* lineage has undergone a significant long term shift in selection pressure and while the causative factors behind this are unknown, the presence of three highly diverged gene families mediating the synthesis of (1,3;1,4)-β-glucan presents an interesting case study in coevolution.

The broad distribution of gene families capable of (1,3;1,4)-β-glucan synthesis across the *CesA* superfamily tree highlights the difficulty in mapping polysaccharide product to phylogenetic structure. This difficulty is compounded by significant systematic confusion; superfamily members in species are named in order of discovery or by homology to different organisms. In Chapter 4, this confusion is addressed using model-based analyses to reconstruct phylogenetic relationships and infer duplication events among the *CesA* and *Csl* genes of 22 fully sequenced angiosperms. The recovered phylogenetic history and identified discriminatory protein motifs were used to construct a revised system for naming new and existing *CesA* and *Csl* genes.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree. I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

……………………………….

Julian George Schwerdt

Acknowledgments

I am grateful to my supervisors Geoff Fincher, Rachel Burton and Maria Hrmova for their guidance and support throughout my candidature. Thanks also to the Australian Centre for Plant Functional Genomics and ARC Centre of Excellence in Plant Cell Walls staff, especially Mahima Krishnan, Neil Shirley, Alan Little, Matt Tucker, Natalie Betts, Helen Collins, David Mathew, Natalie Kibble, Damian Drew, Emma Drew, Andrew Harvey, Kuok Yap, Maria Lombardi and Vincent Bulone. I would also like to acknowledge James Hutton Institute staff Frank Wright, Katrin McKenzie, David Marshall and Robbie Waugh. I would like to thank the Grains Research and Development Corporation, the Australian Research Council and the University of Adelaide for their financial and administrative support. Foremost, I would like to thank my wife, Kate Sanders.

The overall aim of the work described in this thesis was to reconstruct the evolutionary history and dynamics of the enzymes that mediate the synthesis of cell wall polysaccharides in plants, especially the *Poaceae* (grasses). The embryophytes (land plants) have evolved a wide variety of cell wall polysaccharides that differ greatly in their chemistry and ability to perform biological functions in the wall. The embryophytes have simultaneously evolved mechanisms to change the properties of these polysaccharides during growth and development, and in response to abiotic and biotic stress. To understand phylogenies of polysaccharide synthases it is useful consider their evolution in the context of the embryophyte life strategy. These macroevolutionary considerations are outlined below and the chemical diversity of key cell wall polysaccharides is described in relation to current knowledge of the gene families that encode the polysaccharide synthase enzymes.

1.1. Eukaryotes and the Evolution of Land Plants

Molecular phylogenetics places the eukaryotes into six monophyletic protist lineages: Opisthokonta (animals, fungi), Amoebozoa (amoebae, slime moulds), Excavata (oxymonads), Rhizaria, Chromalveolata (ciliates, diatoms, brown algae), and the Archaeplastida, which includes the green algae, red algae, glaucophytes and land plants (Adl et al., 2005). Estimated to have a common ancestor 1,500 Mya (million years ago), the Archaeplastida are distinguished by a carbohydrate- rich cell wall and a chlorophyll *a* or *b* photosynthetic plastid co-opted during a primary endosymbiotic event with a cyanobacterium. Archaeplastida is composed of three distinct lineages: Rhodophyta (red algae), Glaucophyta (microalgae with cyanobacteria-like chloroplasts), and Chloroplastida. Chloroplastida — green algae and land plants — comprises four divisions: Chlorophyta (green algae, e.g. Chlamydomonas), Ulvophyta (green algae, e.g. sea lettuce), Prasinophyta (e.g. Ostreococcus), and Strepsystera (charophyta, embryophytes (land plants)) (Baldauf, 2008).Within Chloroplastida, the evolution of embryophytes (land plants), starting with the colonisation of the land by charophycean green algae, ranks among the most important events in the evolution of life, by significantly altering atmospheric oxygen levels and powerfully influencing the development of terrestrial ecosystems (Scott & Glasspool, 2006). During the early Ordovician period (~443–490 Mya), a lineage of charophycean algae transitioned to a moist, rain-dependent, terrestrial environment. In contrast to the marine algal lineages, it has been suggested that the charophytes were pre- adapted to colonisation because their freshwater habitat was ecologically proximal to the terrestrial environment (Becker & Marin, 2009). Later expansions onto land have occurred at least six times within the green algae, and provide a powerful model for studies in embryophyte physiological adaptation (e.g. Lewis & McCourt, 2004).

The terrestrial charophycean ancestors to the embryophytes most likely entered a niche largely free from similar competitors, although it is probable that cyanobacteria and fungi were present (Labandeira, 2006). Their subsequent adaptive radiation during the Silurian (~419–443 Mya) drove significant morphological change, leading to the enormous diversity in embryophyte body plan and life strategy evident today. By the late Devonian (~360–419 Mya), embryophytes had diversified to produce all of the major extant land plant lineages except the angiosperms (flowering plants) and polypodiaceae ferns. This diversification is estimated to have occurred over only ~46 million years, and thus presents a remarkable adaptative explosion culminating in the appearance of rooted vascular embryophytes during the Paleozoic (Graham et al., 2000).

1.2. Evolution of the Plant Body Plan

In comparison to existing ecosystems, the terrestrial environment was underpopulated when first colonised by the ancestral embryophytes. The subsequent emergence of a broad range of embryophyte body plans and life strategies was most likely driven by a combination of small organismal size (with rapid generations and high mutation rates), founder population effects and selection pressures imposed by a dramatically different colonial environment (Niklas, 2004). There is good evidence that the embryophyte ancestors were mobile with flagella similar to euglenoids, unicellular, and although almost certainly photoautotrophs, probably had the capacity for heterotrophic fallback (Niklas & Kutschera, 2009; Sarkar et al., 2009).

Terrestrial colonisation brought fundamental survival challenges to the first embryophytes, especially in reproduction, which had hitherto relied on an aqueous environment. A possible key adaptation is the development of an alternating sexual cycle between the gametophyte (haploid multicellular phase that produces sperm and egg cells) and the sporophyte (diploid multicellular phase that generates meiospores). The spores produced by the sporophyte adopted an aerial dispersal strategy to cope with desiccation and expanded the genetic pool for sexual recombination. The spermatophyte (seed plant) lineage extended this adaptation with the development of the seed, which can disperse gametes independent of water (Bennici, 2008). Indeed, together with development of the archegonia and antheridia (the organs that produce the sperm and eggs, respectively), the retention of the fertilised egg (hence "embryophyte") (Niklas & Kutschera, 2009) and the alternating multicellular sexual phase, or diplobiontic life cycle, are major distinguishing characteristics of the embryophytes. Aerial dispersal was also a crucial adaptation during the transition from a motile aqueous lifestyle to a stationary terrestrial life strategy, which was possibly a result of, or occurred in conjunction with, the establishment of an autotrophic existence. Land plants subsequently optimised their body plans: to maximise exposure to sunlight for photosynthesis and to the atmosphere for gas exchange; to manage supply and storage of water; to withstand external mechanical forces; and to allow the dispersal of spores for sexual recombination (Sarkar et al., 2008). Without mobility, a flexible, modular body plan provided by the development of branched vegetative tissue was critical to meeting these requirements.

The capacity to branch their cylindrical parenchymatous (versatile ground tissue) bodies into complex architectures of three-dimensional differentiated tissues is another distinguishing characteristic of land plants, and presents an adaptation to address the requirements of stationary life. Branching of aerial parts is enabled and controlled by the shoot (and root) systems, and is driven primarily through the shoot apical meristem that appears in embryogenesis and subsequent meristem activity (Shimizu-Sato & Mori, 2001). The early embryophytes most likely had terminal dichotomous branching patterns (division of the apical cell into two autonomous but morphologically similar branches). The advent of axial branching, where a primordial bud is produced in the apex organogenic zone (Gola, 2014) and can have multiple planes of cell division (cutting faces), enabled architecturally more complex morphologies (Graham et al., 2000; Sussex & Kerk, 2001).

The branching patterns of land plants have been shown to be optimised for lineage-specific environmental conditions (Niklas, 2004; Sussex & Kerk, 2001). Adaptations include flexibility to maximally orient the body towards the sun, and variation in the surface area:volume ratio to either maximise gas exchange or minimise water loss. The early adaptations seen in embryophytes thus prioritised spatially complex and flexible body architectures, in the context of the constraints imposed by autotrophy and immobility. The physical necessity of structural strength to support large complex bodies, the need to defend against pathogen and predation without mobility, and the requirement to control cell expansion, together led to the evolution of a key trait underlying the evolutionary success of embryophytes — their polysaccharide-rich cell wall.

1.3. The Plant Cell Wall and its Evolution

Photoautotrophy had a significant impact during the evolution of the ancestral embryophytes. Photosynthesis brought the capacity to fix carbon, elevating the concentration of carbohydrate solutes within the cell, and consequently increasing osmotic pressure through the uptake of water and hypotonic solutions (Raven et al., 2012). The resulting cell expansion posed significant risks to the cell’s structural viability. Assembling carbohydrates such as glucose into polymers (polysaccharides) reduced osmotic pressure and provided the building blocks for an extra-cellular matrix — the cell wall — capable of withstanding substantial expansion pressures, prohibiting membrane rupture and controlling cell growth (Sarkar et al., 2009). Such are the selective advantages conferred by the cell wall's mechanical and chemical functions that a substantial proportion of embryophyte photosynthetic activity is directed towards its construction (Sørensen et al., 2010). The high tensile strength (stretch resistance) and resistance to compression of the cell wall (Carpita, 1985; Ryden et al., 2003) is critical to supporting the complex branching architectures and enormous body sizes of the embryophytes (Falster & Westoby, 2003), enabling them to contend with predation and adapt their stationary bodies to a varying and highly competitive environment. The cell wall has also played an integral role during the evolution of multicellularity by bonding cells together, controlling cell differentiation, acting as an energy store and mediating chemical signals (Popper et al., 2011).

It is commonly held that the machinery necessary for cell wall synthesis has evolved independently multiple times in prokaryotes, whereas in eukaryotes the cell wall machinery has been acquired via numerous lateral gene transfers during endosymbiotic events (Niklas, 2004; Popper et al., 2011). Such a multifarious evolutionary history is reflected in the diverse cell wall components and structures evident across plant life. Broadly, the definition of a cell wall is a protoplast-created, intra-or extra-cellular matrix attached to a cell membrane (Niklas, 2004). There is a huge diversity of wall constituents including polysaccharides, mucopolysaccharides, peptidoglycans, glycoproteins, glycolipids and lignins. Recent work to systematise cell wall structures across all phyla identified five categories of wall: Type I, plasma membranes; Type II, cell surfaces bound to internal structures; Type III, surfaces with external materials; Type IV, surfaces with vesiculated materials; Type V, surfaces with materials both external and internal to the plasma membrane. All embryophyte cell walls are classed as Type III (Becker, 2000).

Although the rich complexity and heterogeneity across taxonomic and tissue-specific divisions (Burton *et al.*, 2010) of constituent components makes a single encompassing model of the structural organisation of the plant cell wall problematic, some fundamental design principles can be specified. The embryophyte body, being photoautotrophic and stationary, is subject to significant physical forces in order to support its flexible parenchymatous branching architecture, achieve towering heights, and to contend with predation or environmental pressures. Consequently, each cell must resist basic forces of shearing, tension and compression. The plant cell wall is adapted to provide an architecture whose solution is based on the same principles used to engineer fibre-reinforced composite materials, where structural rods or fibres resist the pull of tension and are embedded in a matrix material capable of resisting the squeeze of compression and the sliding cut of shear (Kerstens *et al.*, 2001). The wall of a growing embryophyte cell uses two interconnected polysaccharide networks to provide rigidity and flexibility: the cellulose-hemicellulose network and the pectic polysaccharide matrix. These polysaccharide networks, together with a complex (but relatively sparse) population of structural proteins, comprise the major components of the primary cell wall. The primary cell wall is deposited during tissue growth and characteristically can sustain cell expansion. Secondary cell walls are formed when cells are required to differentiate into specialised cells that acquire further wall polymer components such as lignin (Cosgrove, 2005).

Polysaccharides are the dominant cell wall components. Polysaccharide polymers are primarily constructed from thirteen monosaccharide subunits including the six-carbon pyranose ring conformation of certain hexoses (D-glucose, D-galactose, D-mannose, L-rhamnose and L- fucose), the five-carbon ring furanose form of certain pentoses (L-arabinose, D-xylose and D- apiose), and the acidic sugars (D-galacturonic acid, D-glucoronic acid, L-aceric acid, 3-deoxy-D- mannooctulosonic acid and 3-deoxy-D-lyxo-2 heptulosaric acid) (Doblin et al., 2010). From these building blocks are assembled a diverse array of polysaccharides including cellulose, xyloglucan, (1,3;1,4)-β-glucan, heteroxylans, heteromannans and the group collectively known as pectic polysaccharides (Figure 1).

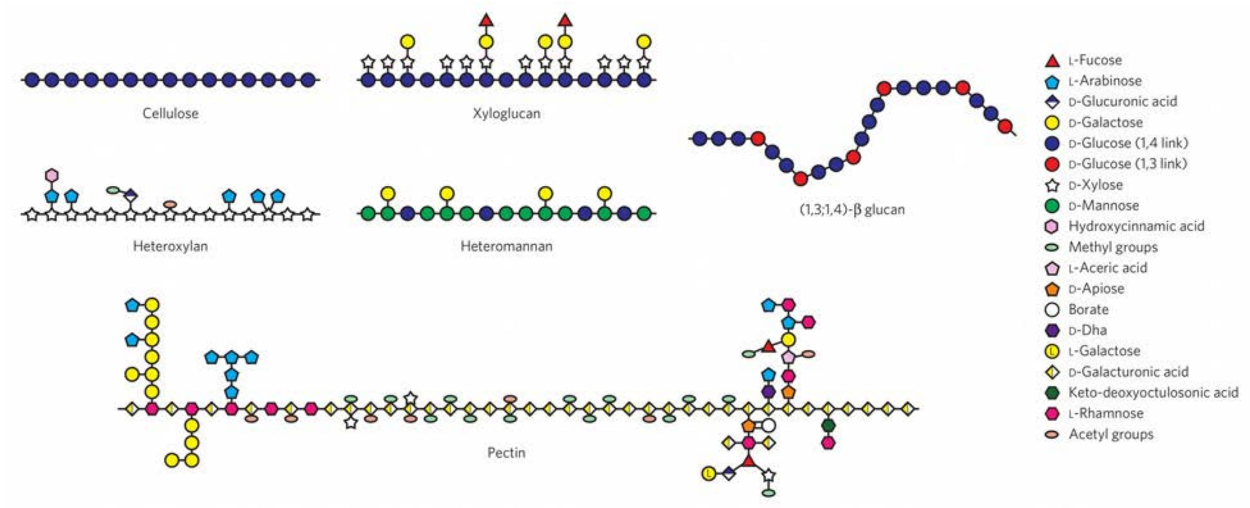


Figure 1: Major embryophyte polysaccharide structures and constituent monosaccharide components. Taken from (Burton et al., 2010).

1.4. Cellulose

The principle load-bearing structures in the cellulose-hemicellulose network are the cellulose microfibrils. Cellulose is the dominant structural polysaccharide in the plant cell wall, accounting for approximately 20–30% of the dry weight of the primary cell wall and 50% of the secondary wall. It is constructed from unbranched (1,4)-β-linked glycosyl residues that are alternately rotated 180° (Figure 1). Cellulose is chemically stable, insoluble and readily forms both intra- and intermolecular hydrogen bonds that, along with van der Waals forces, generate crystalline lateral aggregations of (1,4)-β-linked glucan chains called microfibrils. These extensive hydrogen bonds confer significant strength and rigidity (Brown, 2004). Two major microfibril configurations

(Cellulose I and II) are recognised, although only Cellulose I is found naturally. Cellulose I is characterised by the parallel orientation of the (1,4)-β-linked glucan chains, whereas Cellulose II is an anti-parallel configuration of chains (Brown et al., 1996). Cellulose is synthesised at the plasma membrane from cellulose synthase (CesA) protein complexes. These complexes are spatially organised in symmetric patterns called terminal rosettes, with each subunit synthesising a single chain that assembles with other chains into microfibrils (Brown et al., 1976; Haigler et al., 1980; Somerville, 2006). The precise number of cellulose chains that comprise a microfibril is debated but recent estimates in embryophytes range between 18 and 36 (Newman et al., 2013; Oehme et al., 2015). Spectroscopic observations have measured average microfibril length at 30nm (Caffall & Mohnen, 2009). Spatial limitations caused by turgor pressure pushing the membrane into existing wall material drives newly synthesised microfibrils into an orientation parallel to the cell membrane outer surface. The lamella, a layer one microfibril thick, is progressively deposited to build the primary cell wall in which microfibrils are arranged in an essentially random fashion. In the secondary cell wall, which is often thickened as the need for strength increases in the growing plant, the wall has a more complex lamella organization in which cellulosic microfibrils assume a parallel arrangement (Caffall & Mohnen, 2009). The spatial patterns of the green algae terminal rosettes vary substantially more than the embryophytes. Charophyte species have been shown to contain similar structures to the embryophytes, however other green algae have been recorded to have up to 140 subunits within the terminal rosette and have varied spatial disposition (Tsekos, 1999).

The orientation of the cellulose microfibrils is a critical element in providing mechanical strength to the plant cell wall, with the parallel microfibrils of secondary walls reinforcing the cell wall matrix and generating high tensile strength (Kerstens et al., 2001). The spatial separation between microfibrils is relatively uniform and is thought to be maintained by the hemicellulosic polysaccharides that non-covalently cross-link and enmesh the microfibrils (Verbelen and Vissenberg, 2006).

1.5. Hemicelluloses

The hemicelluloses are a heterogeneous, loosely defined category of polysaccharides characterised by (1,4)-β-linked backbones comprised of glucosyl, mannosyl or xylosyl residues and which excludes pectin or cellulose (Scheller & Ulvskov, 2010). These include xyloglucan, (1,3;1,4)-β-glucan, heteroxylan and heteromannan. The dominant embryophyte hemicellulose is xyloglucan, which accounts for approximately 20–30% of the primary cell wall in eudicots, but is less abundant in coniferophyta, and comprises only ~2–5% of the Poaceae (grasses) wall, where arabinoxylan and (1,3;1,4)-β-glucan are the main hemicelluloses (Hsieh & Harris, 2009).

1.5.1. Xyloglucans

Xyloglucan is a substituted (1,4)-β-linked glucan chain with ~50–75% of the (1,4)-β-glucosyl residues substituted at the C-6 position with α-D-xylosyl whose C-2 can further be substituted with β-D-galactosyl or α-L-fucosyl residues (Figure 1). In the prevailing structural model of the plant cell wall, the hemicelluloses tether and coat cellulose microfibrils, maintaining relatively even spacing between them (Hayashi et al., 1994), although it should be emphasized that there is no evidence to suggest that the 'tethering' is covalent in nature. In xyloglucans, chemical characterisation has identified three operational domains: a microfibril tether-forming domain, a cellulose surface-bound domain, and a domain embedded or nestled between microfibrils (Pauly et al., 1999). This model presents xyloglucan, and tethering hemicelluloses generally, as core determinants of wall strength and extensibility. However, initial proposals that xyloglucan microfibril tether-forming domains perform a load-bearing role have been challenged by evidence suggesting that xyloglucan content does not affect wall tensile strength, and is instead involved in providing extensibility for wall expansion (Whitney et al., 1999; Cavalier et al., 2008).

Precisely how much direct interaction there is between xyloglucan and microfibrils is also complicated by work indicating that only <8% of the cellulose is covered in xyloglucan (Dick- Pérez et al., 2011). The abundance of α-L-fucosyl residues is believed to play a role in determining the binding of the xyloglucan to cellulose microfibrils (Levy et al., 1991), as does the proportion of xylosyl substitutions and the orientation of its glucan backbone. Xyloglucan can adopt a flat orientation that helps it align and bind to the microfibrils (Park & Cosgrove, 2015). Other factors affecting these interactions include the degree of crystallinity of the microfibril, and perhaps also steric hindrance by pectic polysacchardies (Park & Cosgrove, 2015). Further complexity in xyloglucans in the wall is introduced by taxonomic and tissue-specific variation in branching patterns. Poaceae xyloglucan is less highly substituted than that of the eudicots and lacks the α-L-fucosyl residues or α-L-arabinosyl residues that are present rarely in eudicots. These lightly-substituted xyloglucan species are less soluble, which may reflect functional variation in lineages where xyloglucan presence is reduced. Significant taxonomic variation is also found between the bryophytes and angiosperms, with oligosaccharide substituents charged in mosses and liverworts but neutral in the vascular plants (Peña *et al.,* 2008)).

The backbone (1,4)-β-glucosyl residues of xyloglucans can be viewed as cellulosic in nature, but substitutions with mono- or oligosaccharides at C-6 represent an example of how non- cellulosic matrix polysaccharides have evolved varying substituents of otherwise regular polysaccharide chains to sterically limit close molecular alignment and hence aggregation into microfibrils (Burton et al., 2010). Such adaptations have conferred the ability to perform their species, tissue and environmentally specific functional roles as matrix phase and seed storage polysaccharides of the wall (Buckeridge, 2000).

1.*5.2.* Heteroxylans

The relatively low levels of xyloglucan in grass cell walls has led to the proposal that heteroxylans, more specifically arabinoxylan, are to some extent the functional analogue of xyloglucan in the *Poaceae* (Scheller & Ulvskov, 2010). The (1,4)-β-glycan backbones of both xyloglucan and arabinoxylan can adopt a conformation analogous to cellulose (Khnke et al., 2011). The arabinoxylan backbone, like all heteroxylans, is built from (1,4)-β-linked xylosyl residues with diverse substituents that include arabinosyl, galactosyl, glucuronyl, and 4-*O*- methyl glucuronyl residues. In some arabinoxylans, the arabinosyl residues can be further substituted at their C-2 and C-3 positions with phenolic acids such as ferulic acid and p- coumaric acid (Figure 1; Albersheim et al, 2010). Glucuronoarabinoxylan, with higher amounts of glucuronyl and 4-*O*-methyl glucuronyl substitutions, commonly appears in secondary walls. Thus, the heteroxylans are diverse and irregular hemicelluloses with significant variation in substitutions and abundance across taxonomic divisions. For instance, in eudicot secondary walls, large amounts of glucuronoxylans are present, distinguished by prevalent glucuronosyl substitutions, whereas in primary walls of the commelinid monocotyledons (including the *Poaceae*) the abundance of arabinosyl substitutions identifies them as arabinoxylans (Scheller & Ulvskov, 2010).

1.5.3. (1,3;1,4)-β-Glucans

(1,3;1,4)-β-Glucan is a linear, unbranched and unsubstituted chain of β-glucosyl residues polymerised using both (1,3)- and (1,4)-linkages (Figure 1). Strongly associated with the *Poaceae*, they are nonetheless found in other lineages including the Iceland moss (*Cetraria islandical*), certain fungi and horsetail ferns *(Equisetum* sp*.)*

(Fincher, 2009). On average, single (1,3)-linkages are introduced every third or fourth glucosyl residue depending on species and tissues. The (1,3)-linked glucosyl residues confer a higher order of structural diversity on the polysaccharide, which can be considered to be predominantly a co-polymer of (1,3)-linked cellotriosyl and cellotetraosyl oligosaccharides. However, ~10% of the polysaccharide is composed of longer cellodextrins, which have up to 10 or more adjacent (1,4)-glucosyl residues between the single (1,3)-linked residues (Fincher, 2009). These cellotriosyl and cellotetraosyl units are, however, randomly distributed within the polysaccharide. The cumulative effect of two linkage types and randomly arranged oligosaccharide units is a polymer with a multi-level asymmetry that ensures at least partial solubility even at high degrees of polymerisation (DP); the solubility of the polysaccharide can often be predicted from its cellotriosyl:cellotetraosyl ratio, or DP3:DP4 ratio (Fincher, 2009). While the DP3:DP4 ratios of partially soluble (1,3;1,4)-β- glucans from grasses typically lie in the range of 2:1 to 3:1, the corresponding polysaccharides from other taxa display much higher or much lower DP3:DP4 ratios and adopt a less random, less soluble structure that could possibly indicate functional divergence (Fry et al., 2008). There is also evidence to suggest substantial heterogeneity in this fine structural organisation of (1,3;1,4)-β-glucan within the grasses, with (1,3;1,4)-β-glucan from the wheat starchy endosperm wall substantially less soluble than that from either oat or barley (Burton *et al.*, 2009).

The function of (1,3;1,4)-β-glucans in the starchy endosperm of cereal grain cell walls may be two-fold. In addition to its role as a structural component of the wall, (1,3;1,4)-β-glucans appear to act as a significant energy store for the germinating grain, accounting for ~18% of glucose in the grain (Morrall & Briggs, 1978). Functional differentiation of the (1,3;1,4)-β-glucans remains poorly understood although the variation in their fine structures suggests that (1,3;1,4)-β- glucans are employed for diverse purposes. Indeed, they span an exceptionally wide taxonomic grouping (Figure 2), appearing in fungi, Phaeophyceae (brown algae), diatoms, Rhodophyta (red algae), Pteridophytes (ferns, horsetails), Poales (grasses) and bacteria (Popper et al., 2011; Yin et al., 2009). This grouping would place their common ancestor at approximately ~1500 Mya (Wang et al., 1999), yet with the exception of the grasses, (1,3;1,4)-β-glucan is not broadly represented in the major lineages in which it appears. Rather, it appears in just a few species outside the Poales (Bacic et al., 2009). The most phylogenetically parsimonious interpretation of this arrangement implies multiple independent origins for (1,3;1,4)-β-glucan synthesis in eukaryotes and potentially different selection pressure exposure (Burton et al., 2009). The irregular structure of (1,3;1,4)-β-glucans in many grass tissues render it well suited as a matrix polysaccharide, because the irregularly-spaced (1,3)-linkages inhibit close molecular alignment and hence self-aggregation; this in turn increases solubility and forms a flexible porous cell wall matrix phase polysaccharide that is capable of resisting compression and shear forces. An indication of the functional diversity of (1,3;1,4)-β-glucans is the discovery of an endotransglucosylase in *Equisteum* that, *in vitro* at least, uses a (1,3;1,4)-β-glucan donor and xyloglucan acceptor, suggesting a novel interaction between the two polysaccharides. Hrmova et al. (2007) also showed that an endotransglycosylase purified from germinated barley grain could also covalently link oligosaccharides from different wall polysaccharides, including (1,3;1,4)-β-glucans. Homologous genes have been discovered in the Charophytes (Fry et al., 2008).

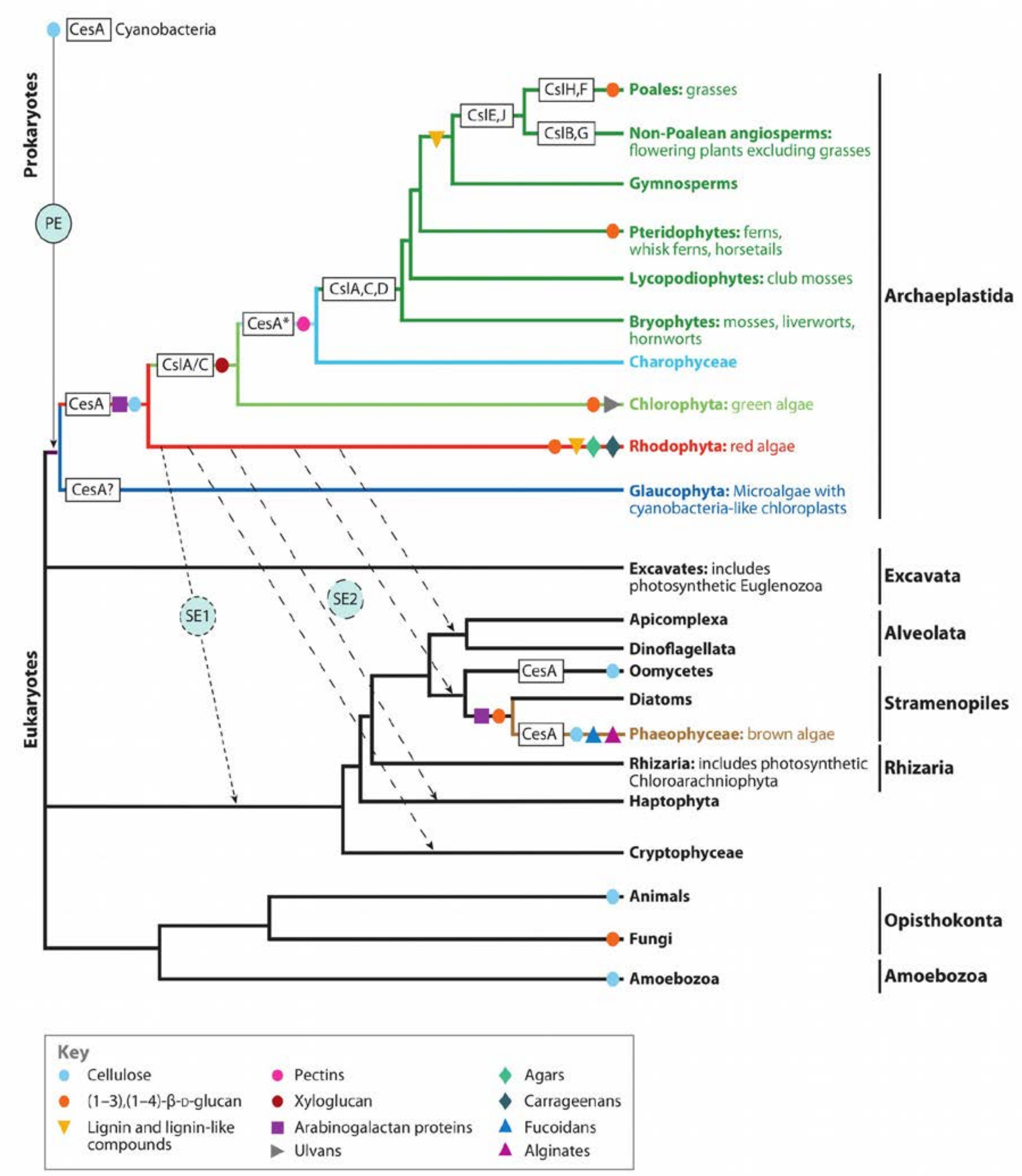


Figure 2: Eukaryote phylogeny detailing the occurrence of major cell wall elements.

Taken from (Popper et al., 2011)

1.5.4. Heteromannans

The heteromannans are comprised of (1,4)-β-linked backbones of D-mannosyl and D-glucosyl residues with α-D-galactosyl substitutions at the C-6 position (Figure 1). This structure is analogous to those of the heteroxylans and the xyloglucans, insofar as they contain a (1,4)-β- glycan backbone substituted with single glycosyl residues that presumably inhibit sterically the alignment and aggregation of these matrix phase polysaccharides.

Heteromannans, especially glucomannans and galactoglucomannans, are present in some legume seeds as storage polysaccharides, including *Ceratonia silique* (carob) and *Amorphophallus konjac* (Buckeridge, Pessoa dos Santos, & Tin, 2000). Other functional roles are possible but have not yet been determined (Scheller & Ulvskov, 2010). The dominant hemicellulose in walls of the Charophytes are heteromannans (Popper, 2003). The polysaccharide is also a major hemicellulose of the gymnosperm secondary wall (~10% w/w of the total wall, relatively rare in the eudicots, a substantial (~15% w/w) component of walls in some ferns (e.g. *Pteridium*) and abundant in bryophytes and lycophytes (Popper, 2003). Such a distribution has prompted speculation that other hemicelluloses, in particular the xyloglucans and heteroxylans, replaced the heteromannans in later diverged plant lineages such as the spermatophytes (Scheller & Ulvskov, 2010).

1.6. Pectin

The pectin network, generally considered to be an independent organisation of polysaccharides around the cellulose-hemicellulose framework, nevertheless is interconnected, possibly covalently, with cellulose and hemicelluloses like xyloglucan (Mohnen, 2008; Dick-Pérez et al., 2011). Pectins are highly complex polymers that are critical components in many, especially primary, plant cell walls, forming ~30% of the eudicot primary wall and ~10% dry weight in primary walls of the Poales. The pectic polysaccharides are polymers of D-galacturonyl residues and include homogalacturonan, xylogalacturonan, apiogalacturonan, rhamnogalacturonan I and rhamnogalacturonan II. The dominant pectic polysaccharide is homogalacturonan, accounting for over 60% dry weight of the pectin in some taxa (Caffall & Mohnen, 2009). Homogalacturonan is a linear polymer of (1,4)-α-linked-D-galacturonyl residues, and can be heavily methyl-esterified to more than 80% by some observations (Willats et al., 2001). Unmethylated galacturonic acid residues can form calcium bridges with adjacent unmethylated regions from other homogalacturonan molecules and other pectic polysaccharides, assembling into a stable gel matrix (Caffall & Mohnen, 2009). Rhamnogalacturonan I is a heterogenous polysaccharide that has a repeating rhamnosyl and galacturonosyl disaccharide backbone with variable side chain residues of arabinosyl, galactosyl, fucosyl and glucuronyl residues. Rhamnogalacturonan II is a highly complex polysaccharide composed of twelve monosaccharides and is profusely branched and contains chelated borate ions; a galacturonan backbone is substituted with four side chains classified (A–D) (Caffall & Mohnen, 2009).

Along with homogalacturonan calcium bridges, pectins are likely to be interconnected through glycosidic linkages between homogalacturonan and rhamnogalacturonan I and II, borate ester crosslinks and covalent links to phenolic compounds (Vincken et al., 2003). Evidence suggests that pectin is a highly dynamic polysaccharide network displaying much functional plasticity, including the capacity to adopt a gelled or more rigid conformation, through varying levels of methyl esterification (Palin & Geitmann, 2012). Pectins have been shown to bind non-covalently to cellulose probably through their neutral side chains, such as those with higher arabinosyl and galactosyl substitutions (Zykwinska et al., 2005). Observations that cellulose microfibrils become more rigid and tightly packed after pectin removal are consistent with a central role of pectins in cell growth and elongation. Pectin involvement in separating cellulose microfibrils also plays a role in influencing porosity of the wall and thus regulating the transport of signals, proteins and other polysaccharides (Dick-Prez et al., 2012).

The diverse linkages and residues that comprise the pectic polysaccharides impart an ability to fine tune the wall network for tissue- and environmentally-specific properties (Vincken et al., 2003). Such structural flexibility is interesting in light of the basal position of pectin in the eukaryote phylogeny. Current sampling implies a single origin for pectin synthesis appearing with the common ancestor of embryophytes and charophytes (Figure 2);(Popper et al., 2011). That this archaeplastida lineage is also the most morphologically complex is an intriguing correlation.

1.7. Plant Cell Wall Biosynthesis

Ultimately, photosynthesis is the source of all carbohydrates used to assemble plant cell wall polysaccharides. The first products are a collection of hexose monophosphates produced by phosphoglycomutases and 6-phosphate isomerases (Albersheim et al, 2010). These sugars are converted into nucleoside diphosphate sugars (NDP-sugars), through either pyrophosphorylase activity or latter interconversions, such as membrane-bound sucrose synthases. Only after their activation to nucleotide sugars can the sugar residues be assembled into cell wall polysaccharides. (Gibeaut, 2000; Seifert, 2004; Bar-Peled & O'Neill, 2011). The enzymes that catalyse the transfer of NDP-sugar donors onto various acceptors (such as plant cell wall polysaccharide chains) through glycosidic linkages are called glycosyltransferases.

1.7.1. Glycosyltransferases

Glycosyltransferases (GT) are a very large group of membrane bound proteins that are estimated to represent ~1% of open reading frames in sequenced genomes (Coutinho *et al.*, 2003). The inherent complexity of GT biochemical characterisation made classification using the enzyme commission (EC) framework difficult (http://www.chem.qmul.ac.uk/iubmb/enzyme/). Current nomenclature is determined by amino acid similarity and GTs have been resolved into 97 identified families, which are curated in the Carbohydrate-Active Enzymes database (CAZy; http://www.cazy.org; Coutinho et al., 2013) along with enzyme families dedicated to carbohydrate degradation, including glycoside hydrolases (GH), polysaccharide lyases (PL), carbohydrate esterases (CE) and associated enzymes designated as having auxiliary activities (AA). The classification system is built using clusters of hydrophobic sequences assigned by sequence similarity to a biochemically characterised enzyme and is modular such that one protein may belong to multiple families (Lombard *et al.*, 2014).

The general definition of protein in the GT family is an enzyme that uses an activated donor sugar substrate that has a phosphate leaving group; nucleotide sugars are the most important of these activated sugar donors. The acceptor substrate is usually another sugar molecule, which is generally at the non-reducing end of the nascent polysaccharide but proteins, lipids, nucleic acids and other small molecules are occasionally used (Lairson *et al.,* 2008).

GTs are defined by two three-dimensional (3D) protein folds, GT-A and GT-B, that split the family into two functional-structural categories (Lairson *et al.*, 2008). The GT-A fold comprises a Rossmann fold-like sandwich of a seven stranded β-sheet between α-helices. Two additional conserved domains are recognised, one encompassing an NH2- terminal domain ending in a conserved DxD motif (where x is any residue), the other with a region containing acceptor sites positioned near an interface of a secondary β-sheet and the region around β-strand 6. GT-B comprises two Rossman folds with the catalytic site positioned between and connected by, a linking domain. Whilst the COOH-terminal region shows high levels of conservation, the helices and loops that reach into the active site show significant variation, indicating their utilisation of diverse acceptors (Breton et al., 2006).

Many GTs utilise identical donor or acceptor substrates but there is little sequence homology between the majority of GT families (Breton et al., 2001), suggesting a complex evolutionary history with separate protein lineages converging on GT functionality. However, the discovery in archaea of a single GT family for both GT-A (GT2) and GT-B folds (GT4) indicates a single evolutionary origin for these proteins. Given that archaea diverged from eukaryotes and bacteria between 2700 and 4100 Mya (Gribaldo et al., 2006), the ancestral GT must be ancient with vastly diverged extant descendants. Complicating this is a recently proposed GT-C fold that was predicted using BLAST (Altschul et al, 1990) and demonstrated by an oligosaccharyl-transferase crystal structure from the archaea *Pyrococcus furiosus* (Igura et al., 2008). However, the structure revealed that the GT-C domain corresponds mainly to transmembrane regions and not catalytic sites, which casts some doubt on the credibility of GT-C as a classifying fold. There are ~800 experimentally determined glycosyltransferase 3D protein structures available at the Protein Data Bank (PDB; http://www.rcsb.org). These structures span the GT families 1, 2, 5, 6, 8, 9, 13, 15, 20, 27, 28, 42 and 43 across prokaryotes such as *Agrobacterium tumefaciens, Escherichia coli* and *Bacillus subtilis*, eukaryotes including human, cow and mouse, a viral phage T4, but as yet no embryophytes (Christelle Breton et al., 2006). However, a major step towards a structural account of embryophyte polysaccharide synthases was achieved with the crystal structure of *Rhodobacter sphaeroides* GT2 bacterial cellulose synthase A and B subunits (BcsA, BcsB) (Morgan *et al.*, 2013; Morton *et al*., 2014). The BcsA structure revealed four NH2-terminal and four COOH-terminal transmembrane helices divided by a cytosolic domain that includes the GT-A fold, which comprises a seven stranded β-sheet and seven α-helices connected to the transmembrane region by three amphipathic interface helices (IF) (Morgan et al., 2013). A pore is formed by six transmembrane helices (TM3–8) where the (1,4)-β-glucan chain translocates and emerges close to the interface with BcsB and into the periplasm. The highly conserved GT2 motif Q/RxxRW is located on the intracellular amphipathic helix IF2 and is shown to be the binding site for the disaccharide acceptor. Another characteristic and highly conserved GT2 motif, TED, is shown to be on a 'finger' helix responsible for translocating the (1,4)-β-glucan chain into the pore by one glucose unit (Morgan *et al.*, 2014).

1.8. The Cellulose Synthase Superfamily

Embryophyte homologues to bacterial cellulose synthases were initially identified in *Gossypium hirsutum* (cotton) and *Oryza sativa* (rice) cDNA libraries by sequence similarity, particularly by the presence of the core catalytic Q/RxxRW motif in all representatives. The two initial cellulose synthase cotton genes, designated *CesA1* and *CesA2,* were shown to bind UDP-D-glucose and were highly expressed in developing cotton fibre tissue during secondary cell wall cellulose synthesis (Pear et al., 1996; Arioli et al., 1998). It was recognized through analyses of EST libraries that the cellulose synthase *CesA* genes were but one clade in a much larger group of genes. More recently, the publication of complete embryophyte genome sequences has confirmed this large and diverse lineage of homologues that in most plants contains about 50 genes and which is now commonly known as the cellulose synthase (*CesA*) gene superfamily (Richmond and Somerville, 2000; Hazen et al., 2002).

The GT2 cellulose synthase superfamily was initially resolved into the *CesA* and six *cellulose synthase-like* (*Csl*) clades: *CslA*, *CslB*, *CslC*, *CslD*, *CslE* and *CslG* (Richmond and Somerville, 2000)*.* Subsequent work, especially in *Poaceae*, revealed three additional lineages: *CslF* (Hazen et al., 2002), *CslH* (Hazen et al., 2002) and *CslJ* (Farrokhi et al., 2006; Fincher, 2009). The superfamily spans Archaeplastida but is most abundant in Strepsystera, with Chlorophyta containing ~1 and embryophyta ~40–60 (Yin et al., 2009) members. As with the evolution of photosynthesis, it is hypothesised that the Archaeplastida cellulose synthase superfamily originated by endosymbiotic transfer from cyanobacteria. Cellulose biosynthesis has been demonstrated in two cyanobacteria, *Anabaena* sp*. a*nd *Nostoc punctiforme,* with putative *CesA* genes having homology with domains then considered specific to eukaryotes (Nobles *et al.*, 2001). In a subsequent phylogeny of prokaryote and eukaryote GT2s, a marine cyanobacterium (*Synechoccus* sp.) was shown to contain a *CesA* lineage monophyletic to the embryophyte *CesA* and *CslB*/*D*/*E*/*F*/*G* clades (Nobles & Brown, 2004). Notably however, *CslA* and *CslC* comprised a sister group to the other embryophyte *CesA/Csl* genes and *Synechoccus CesA1.* This implied that *CslA* and *CslC* genes represent an independent lineage to *CesA* and *CslB/D/E/F/G/H/J*, possibly originating from a separate endosymbiotic transfer event

(Figure 2).

As the taxonomic breadth of cellulose synthase superfamily sampling widened, it became evident that substantial diversification has occurred since the emergence of embryophytes, especially within the *Poaceae*, where three major grass-specific lineages have been identified, namely *CslF, CslJ* and *CslH*. However, the systematics of the superfamily is far from resolved, including for instance whether *CslJ* is found only in *Poaceae* (Fincher, 2009; Yin et al., 2009). Functionally, the generally accepted view is that the *CesA* family synthesises cellulose whereas the *Csl* genes are involved in biosynthesis of hemicellulose polysaccharides (Richmond & Somerville, 2000).

1.8.1. *CesA*

The *CesA* family was the first identified and is the most well-characterised lineage in the cellulose synthase superfamily. CesA proteins contain approximately 900–1000 amino acid residues and have an approximate molecular weight of 100 kDa (Atanassov et al., 2009). Intra- species sequence similarity is relatively high (64% in *Arabidopsis*) and intron positions are conserved across Archaeplastida (Roberts & Roberts, 2004). CesA proteins are predicted to contain eight transmembrane helices (TMH1–8) that span the cytosolic catalytic region between TMH2 (near the NH2-terminus) and TMH3 (near the COOH-terminus). This cytosolic domain contains the characteristic GT2 motifs D,D,D and Q/RxxRW. It is highly conserved across embryophytes with the exception of a small divergent section that was initially considered to be a hypervariable region, but with wider taxonomic sampling was found to vary according to species and was designated the class-specific region (CSR) (Somerville, 2006). The alignment of bacterial and plant CesA sequences revealed a second CSR located at the NH_2-terminus with both appearing to be embryophyte-specific inserts (Pear et al., 1996). The NH_2-terminus also contains phosphorylation sites and putative RING-finger domains; the latter are zinc (Zn) binding domains that have been shown to be involved in protein-protein interactions. In cotton fibres, *GhCesA1* has been demonstrated to bind Zn^2+ and, along with *GhCesA2*, forms homo- and heterodimers (Kurek et al., 2002).

Evidence for CesA dimerisation reinforced the rosette, or terminal complex (TC), model of cellulose microfibril synthesis. The prevailing model depicts the membrane bound TC as a hexameric association of CesA complexes (CSC) with each CSC constructed from six individual CesA proteins (Brown et al., 1976; Haigler et al., 1980; Doblin et al., 2002). Two orthologous groups of three *CesA* genes encode the CSC proteins. In *Arabidopsis*, *AtCesA1, AtCesA3* and *AtCesA6* were shown to be necessary for cellulose synthesis at the stage when primary walls were forming (Eckardt, 2003) and secondary wall CSCs have been associated with *AtCesA4, AtCesA7* and *AtCesA8* (Gardiner et al., 2003). Barley transcript studies similarly revealed that *HvCesA1, HvCesA2* and *HvCesA6* formed a co-expressed group transcribed in tissue associated with primary wall cellulose synthesis. In maturing root and stem, where secondary wall deposition is occurring, *HvCesA4, HvCesA7* and *HvCesA8* were co-expressed. The

Arabidopsis primary and secondary wall CSC genes form homologous clusters with the barley co-expressed groups (Burton et al., 2004), which implies that the functional division between primary and secondary cell wall *CesA* genes is at least as old as the 140–150 Myr eudicot- monocot split (Chaw et al, 2004).

1.8.2. *CslD* and *CslF*

The *CslD* and *CslF* families comprise a sister clade to the *CesA* genes (Farrokhi et al., 2006; Yin et al., 2009). The *CslF* genes were first identified following the completion of the rice (*Oryza sativa*) genome sequence (Hazen *et al.*, 2002). Phylogeny and comparative sequence analysis to *Arabidopsis* suggested *CslF* to be a cereal-specific family and currently has only been sampled in the Poaceae (Yin et al., 2009). Seven genes were initially identified (*CslF1–7*), although copy numbers vary across species with barley now shown to have ten *CslF* members (Burton et al., 2008; Schreiber et al., 2014). Eight transmembrane helices are predicted; these have an intra-species amino acid identity of ~40–65% and share *CesA* CSR and catalytic motifs. Intron positions are also shown to be relatively conserved. CslF enzymes range from ~810–947 amino acids in length and so are slightly truncated when compared with the *CesA* protein sequences (Burton et al., 2008).

The *CslF* genes were shown to perform a central role in (1,3;1,4)-β-glucan biosynthesis when (1,3;1,4)-β-glucan, not present in eudicots, was detected in the walls of transgenic *Arabidopsis* expressing rice *CslF* proteins (Burton et al., 2006). Six of the rice genes were shown to cluster across 118 kilobases and syntenic regions containing similar clusters of *CslF* genes are present in all *Poaceae* so far examined, including barley (where the cluster overlaps a (1,3;1,4)-β-glucan quantitative trait locus (QTL)), *Sorghum bicolor* and *Brachypodium distachyon* (Burton et al., 2006; Ermawar et al., 2015; Fincher, 2009). Based on transcript abundance, *CslF6* is the major gene involved in (1,3;1,4)-β-glucan synthesis in reproductive and vegetative tissues and is one of the few family members not in the *CslF* gene cluster. *CslF6* orthologues are also notable for a conserved, but unique to *CslF6*, ~55 amino acid insert that is predicted to adopt a loop conformation but whose precise role remains unknown (Fincher, 2009). The CslF6 enzyme, like other cellulose synthase superfamily proteins is membrane bound and was recently demonstrated to have a final location in the plasma membrane. This is counter to the commonly held assumption that matrix polysaccharides are synthesised in the Golgi. Such observations are integral to ongoing discussions about the model and assembly of (1,3;1,4)-β-glucan and other matrix phase polysaccharides (Burton et al., 2010; Kim et al, 2015; Wilson et al., 2015).

As with the *CesA* genes, some *CslD* sequences contain putative RING-finger domains and are the family most homologous to the *CesA* genes (Richmond & Somerville, 2000). *CslDs* are relatively conserved across intra- and inter-species divisions, with amino acid identities ranging between 64–91%, they contain eight putative transmembrane helices and have a homologous CesA CSR (Doblin et al., 2001). A definite function has not yet been assigned to *CslD,* but analysis of *Arabidopsis* mutant phenotypes has suggested a role in mannan biosynthesis. However, activity could only be demonstrated when two genes, *AtCslD2* and *AtCslD3,* were co- expressed, which suggested CslD enzymes might be organised in a protein complex, potentially mediated by the RING-finger domains (Verhertbruggen et al., 2011; Yin et al., 2009).

Recent work identifying barley *CslD* and *CslA* as potential core genes in resistance to powdery mildew pathogens also supports a role in mannan biosynthesis, as *CslA* genes are also thought to make mannans, however the role of mannan in pathogen response is unclear (Douchkov et al., 2014) with potential factors such as wall strengthening yet to be assessed. Mannan abundance in early diverging embryophytes, and the suggestion that mannan was replaced by other hemicelluloses in later diverged plants, is sensible in light of the basal position of *CslD* on the embryophyte tree (Popper et al., 2011). *CslD* groups with *CslA* and *CslC,* representing the only *Csl* members in the bryophytes (Figure 2)*;*(Popper et al., 2011; Yin et al., 2009).

1.8.3. *CslA* and *CslC*

CslA and *CslC* are ancient lineages found in both Chlorophyta (green algae) and Strepsystera (Figure 2)*;*(Liepman & Cavalier, 2012; Popper et al., 2011). Their protein domain organisation is consistent with the hypothesis that they originated from a cyanobacterial endosymbiotic transfer separate to that of the *CesA* genes and did not diverge from a common ancestor in embryophytes. CslA and CslC contain four or five putative TMHs in contrast to the eight present in other CesA superfamily lineages, suggesting a potentially different pore structure from which the polysaccharide chain emerges. With approximately 500 amino acid residues, the CslA and CslC enzymes are approximately half the length of CesA proteins and are the most highly conserved clade (Yin et al., 2009). Additionally, CslA and CslC differ from the other lineages in having a highly basic loop in place of a putative TMH and an acidic loop between TMH3 and TMH4 within range of the pore (Brown and Saxena, 2007).

Expression in *Drosophila* has demonstrated that certain *Populus trichocarpa* (poplar) *CslA* genes (particularly *PtCslA1* and *PtCslA3*) encode mannan or glucomannan synthases (Liepman et al., 2007; Suzuki *et al.*, 2006). Heterologous expression also implied that CslA is completely functional by itself, unlike proteins in the CSC (Dhugga, 2012). Further functional evidence includes the identification of the *Amorphophallus konjac* (voodoo lily) homologue of *Arabidopsis CslA3* as part of the mannan synthesis pathway (Gille et al., 2011). *Amorphophallus* sp. has a specialised storage tissue called the corm where very large deposits of glucomannan are present (Gille et al., 2011). *Arabidopsis CslA2, CslA3 CslA7,* and *CslA9* have also been implicated in stem strength and embryogenesis (Goubet et al., 2009).

Analysis of differentially expressed seed transcripts implicated the *Tropaeolum majus* (nasturtium) homologue of *Arabidopsis CslC4* in the synthesis of xyloglucan, the dominant nasturtium storage polysaccharide (Cocuron et al., 2007). Heterologous expression of *TmCsl* and *AtCslC4* genes synthesised a (1,4)-β-linked glucan with a low degree of polymerization (DP). However, attempts to build side chain substitution through the co-expression of *CslC4* and a xyloglucan xylosyltransferase (*AtXXT1*) did not produce xyloglucan, although it did alter the DP of the oligoglucoside chain (Cocuron et al., 2007). Characterisation of *CslC* genes in barley revealed similar patterns of transcription where *HvCslC3* is co-expressed with *HvGT3*, a gene with strong homology to *AtXXT1*. Phylogenetic analysis grouped *HvCslC3* with *AtCslC4* further suggesting *CslC* involvement in xyloglucan backbone assembly (Dwivany et al., 2009). However, another barley gene *HvCslC2* did not co-express with *HvGT3* or a xyloglucan endotransglycosylase and it must be noted that not all genes in a Csl clade necessarily mediate in the synthesis of the same polysaccharide. Labeling experiments with a polyclonal antibody showed *HvCslC2* residing in the plasma membrane and not in the Golgi where xyloglucan synthesis has been demonstrated, and this led to the proposal that *HvCslC2* might instead be involved in cellulose or callose synthesis (Dwivany et al., 2009). The functional division between xyloglucan and cellulose or callose synthesis corresponds with a phylogenetic division that includes embryophytes, bryophytes and lycophytes, indicating a potential, ancient divergence (Dwivany et al., 2009; Liepman & Cavalier, 2012).

1.8.4. *CslB* / *CslH* and *CslE* / *CslG* / *CslJ*

The remaining *Csl* families, *CslB*/*H*/*E*/*G*/*J,* comprise a sister clade to *CesA, CslD* and *CslF* and are restricted to the embryophytes. However, homologous sequence fragments have been identified in bryophytes, hinting at ancient gene loss events (Yin et al., 2009). Although less well- characterised, CslB/H/E/G/J proteins are structurally similar to CesA/CslD/F, containing eight TMHs and lengths of ~700 amino acids. *CslB* and *CslH* group together in the *Csl* phylogeny and are eudicot- and monocot-specific, respectively. That they are monophyletic implies that a single copy was present in the eudicot-monocot ancestor and subsequently underwent substantial duplication events. Indeed, many of the sampled species contain paralogous and monophyletic gene clusters that indicate recent post-speciation duplication (Hamann et al., 2004; Yin et al., 2009).

In barley, *CslH* has been demonstrated to mediate the synthesis of (1,3;1,4)-β-glucan in transgenic *Arabidopsis* plants (Doblin et al., 2009). Barley has a single *CslH* which is 62-69% identical to the paralogous cluster of three rice *CslH* genes. Notably, *HvCslH* expression is not coordinated with the *HvCslF* genes and so potentially represents a second independent mechanism of (1,3;1,4)-β-glucan biosynthesis in barley (Doblin et al., 2009). That *CslF* and *CslH* are from two highly diverged gene lineages also supports a case for the convergent evolution of (1,3;1,4)-β-glucan biosynthesis in Poaceae (Figure 2)*;*(Fincher, 2009).

The eudicot specific *CslB* clade has not yet been functionally characterised. As there is no observed (1,3;1,4)-β-glucan in eudicots, *CslB* and *CslH* present an intriguing functional divergence for such closely related families. In *Arabidopsis,* the *CslB* family comprises a clustered paralogous group of six genes, which are the least expressed of all the *Arabidopsis Csl* genes and have been speculated to be involved in cell expansion (Hamann et al., 2004).

CslE, CslG and *CslJ* genes comprise the sister lineage to *CslB* and *CslH.* The *CslE* and *CslG* genes are distinguished by monophyletic paralogous gene clusters. This pattern does not extend to *CslJ* which has a different, and less conserved, intron structure to the eudicot-specific *CslG,* its closest relative (Yin et al., 2009). *CslE* genes are expressed in a broad range of tissues and in rice are reported to have an LxxRW sequence in the Q/RxxW domain, but despite this intriguing difference, the functions of these genes remain uncharacterised (Hazen et al.,2002). It has been suggested that *CslJ* is implicated in (1,3;1,4)-β-glucan biosynthesis (Farrokhi et al., 2006; Fincher, 2009), but functional data on this entire clade has not yet been published.

1.9. Aims of the Current Studies

The *CesA* superfamily is central to the biosynthesis of archaeplastida cell wall carbohydrates. Meaningful classification of these important genes is impeded by extreme challenges in characterising their diverse functional roles. However, there is significant potential for computational methods to complement molecular and biochemical experimental work. For example, phylogenetic reconstruction of the *CesA* superfamily is needed to provide an evolutionary context to functional studies. The overall aim of the work described in this thesis was to provide the first large scale analysis of the evolutionary history and dynamics of the *CesA* superfamily in fully sequenced higher plants. Within this overall aim, more specific objectives included; 1. testing hypotheses of selection dynamics of important gene families using a broad representation of embryophyte species; 2. contributing to the understanding of *CesA* superfamily evolution in the major embryophyte divisions; 3. identifying regions of the encoded enzymes that are subject to intense natural selection pressure in specific lineages and 4. proposing a unifying gene nomenclature protocol that can be used for *CesA* and *Csl* genes across the embryophytes based on phylogenetic relationships.

1.10. References

Adl, S. M., Simpson, A. G. B., Farmer, M. A, Andersen, R. A, Anderson, O. R., Barta, J.R., Taylor, M. F. J. R. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *The Journal of Eukaryotic Microbiology*, *52*(5), 399–451. http://doi.org/10.1111/j.1550-7408.2005.00053.x

Arioli, T., Peng, L., Betzner, A. S., Burn, J., Wittke, W., Herth, W., Camilleri, C., Hofte, H., Plazinski, J., Birch, R., Cork, A., Glover, J., Redmond, J., & Williamson, R.E. (1998). Molecular analysis of cellulose biosynthesis in Arabidopsis. *Science*, *279*(5351), 717-720. http://doi.org/10.1126/science.279.5351.717

Atanassov, I. I., Pittman, J. K., & Turner, S. R. (2009). Elucidating the mechanisms of assembly and subunit interaction of the cellulose synthase complex of Arabidopsis secondary cell walls. *Journal of Biological Chemistry*, *284*(6), 3833–3841. http://doi.org/10.1074/jbc.M807456200

Bacic A, Fincher G, Stone B (2009) Chemistry, biochemistry, and biology of (1-3)-β -glucans and related polysaccharides. Academic Press/Elsevier, Amsterdam.

Baldauf, S. L. (2008). An overview of the phylogeny and diversity of eukaryotes, *46*(3), 263–273. http://doi.org/10.3724/SP.J.1002.2008.08060

Bar-Peled, M., & O'Neill, M. A. (2011). Plant nucleotide sugar formation, interconversion, and salvage by sugar recycling. *Annual review of plant biology*, *62*, 127-155. http://doi.org/10.1146/annurev-arplant-042110-103918

Becker, B., & Marin, B. (2009). Streptophyte algae and the origin of embryophytes. *Annals of Botany*, *103*(7), 999–1004. http://doi.org/10.1093/aob/mcp044

Bennici, A. (2008). Origin and early evolution of land plants: Problems and considerations. *Communicative & Integrative Biology*, *1*(2), 212–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686025&tool=pmcentrez&rendertype=abstract

Breton, C., Mucha, J., & Jeanneau, C. (2001). Structural and functional features of glycosyltransferases. *Topology*, *83*.

Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J., & Imberty, A. (2006). Structures and mechanisms of glycosyltransferases. *Glycobiology*, *16*(2), 29–37. http://doi.org/10.1093/glycob/cwj016

Brown, R.M., Saxena, I.M & Kudlicka, K. (2004). Cellulose Biosynthesis in higher plants. *Trends in Plant Science, 5*(1), 149–156.

Brown, R.M., Willison, M & Richardson, C.L. (1976). Cellulose biosynthesis in *Acetobacter xylinum*: Visualization of the site of synthesis and direct meansurement of the *in vivo* process. *Proceedings of the National Academy of Sciences of the United States of America. 73(12), 4565–4569.* http://dx.doi.org/10.1073/pnas.73.12.4565

Brown, R. M. (2004). Cellulose Structure and Biosynthesis : what is in store for the 21th century? *Journal of Polymer Science*, *42*(February), 487–495.

Buckeridge, M. S., Pessoa dos Santos, H., & Tiné, M.S. (2000). Mobilisation of storage cell wall polysaccharides in seeds. *Plant Physiology and Biochemistry*, *38*(1-2), 141–156. http://doi.org/10.1016/S0981-9428(00)00162-5

Burton, R. A, Gidley, M. J., & Fincher, G. B. (2010). Heterogeneity in the chemistry, structure and function of plant cell walls. *Nature Chemical Biology*, *6*(10), 724–732. http://doi.org/10.1038/nchembio.439

Burton, R. A, Jobling, S. A, Harvey, A. J., Shirley, N. J., Mather, D. E., Bacic, A., & Fincher, G. B. (2008). The genetics and transcriptional profiles of the cellulose synthase-like HvCslF gene family in barley. *Plant Physiology*, *146*(4), 1821–1833. http://doi.org/10.1104/pp.107.114694

Burton, R. A, Shirley, N. J., King, B. J., Harvey, A. J., & Fincher, G. B. (2004). The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant Physiology*, *134*(1), 224–236. http://doi.org/10.1104/pp.103.032904

Burton, R.A., Wilson, S. M., Hrmova, M., Harvey, A. J., Shirley, N. J., Medhurst, A., Fincher,G. B. (2006). Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1,3;1,4)-beta-D-glucans. *Science (New York, N.Y.)*, *311*(5769), 1940–2. http://doi.org/10.1126/science.1122975

Burton, R. A., & Fincher, G. B. (2009). (1,3;1,4)-β-D-glucans in cell walls of the Poaceaeee, lower plants, and fungi: A tale of two linkages. *Molecular Plant*, *2*(5), 873–882. http://doi.org/10.1093/mp/ssp063

Caffall, K. H., & Mohnen, D. (2009). The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydrate Research*, *344*(14), 1879–1900. http://doi.org/10.1016/j.carres.2009.05.021

Carpita, N. C. (1985). Tensile strength of cell walls of living cells. *Plant Physiology*, *79*(2), 485– 488. http://doi.org/10.1104/pp.79.2.485

Cavalier, D.M., Lerouxel, O., Neumetzler, L., Yamauchi, K., Reinecke, A., Freshour, G., Zabotina, O.A., Hahn, M.G., Burgert, I., Pauly, M., Raikhel, N.V., & Keegstra, K. (2008). Disrupting Two *Arabidopsis thaliana* Xylosyltransferase Genes Results in Plants Deficient in Xyloglucan, a Major Primary Cell Wall Component. *Plant Physiology*, *20*(6), 1519–1537. http://doi.org/10.1105/tpc.108.059873.

Chaw, S. M., Chang, C. C., Chen, H. L., & Li, W. H. (2004). Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *Journal of Molecular Evolution*, *58*(4), 424–441. http://doi.org/10.1007/s00239-003-2564-9

Cocuron, J.-C., Lerouxel, O., Drakakaki, G., Alonso, A. P., Liepman, A. H., Keegstra, K., Wilkerson, C. G. (2007). A gene from the cellulose synthase-like C family encodes a beta- 1,4 glucan synthase. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(20), 8550–8555. http://doi.org/10.1073/pnas.0703133104

Coutinho, P. M., Deleury, E., Davies, G. J., & Henrissat, B. (2003). An evolving hierarchical family classification for glycosyltransferases. *Journal of Molecular Biology*, *328*(2), 307– 317. http://doi.org/10.1016/S0022-2836(03)00307-3

Cosgrove, D.J. (2005). Growth of the plant cell wall. *Nature Reviews of Molecular Cell Biology*, *6,* 850-861.

Dhugga, K. S. (2012). Biosynthesis of non-cellulosic polysaccharides of plant cell walls. *Phytochemistry*, *74*, 8–19. http://doi.org/10.1016/j.phytochem.2011.10.003

Dick-Perez, M., Wang, T., Salazar, A., Zabotina, O.A., & Hong, M. (2012). Multidimensional solid-state NMR studies of the structure and dynamics of pectic polysaccharides in uniformly ^{13}C-labeled Arabidopsis primary cell walls. *Magnetic Resonance in Chemistry*, *50*(8), 539–550. http://doi.org/10.1002/mrc.3836

Dick-Pérez, M., Zhang, Y., Hayes, J., Salazar, A., Zabotina, O.A, & Hong, M. (2011). Structure and interactions of plant cell-wall polysaccharides by two- and three-dimensional magic- angle-spinning solid-state NMR. *Biochemistry*, *50*(6), 989–1000. http://doi.org/10.1021/bi101795q

Doblin, M.S., Kurek, I., Jacob-Wilk, D & Delmer, D.P. (2002). Cellulose Biosynthesis in Plants: from Genes to Rosettes. *Plant Cell Physiology, 43*(12), 1407-1420. http://dx.doi.org/10.1093/pcp/pcf164

Doblin, M. S., De Melis, L., Newbigin, E., Bacic, A, & Read, S. M. (2001). Pollen tubes of *Nicotiana alata* express two genes from different beta-glucan synthase families. *Plant Physiology*, *125*(4), 2040–2052. http://doi.org/10.1104/pp.125.4.2040

Doblin, M. S., Pettolino, F. A., Wilson, S. M., Campbell, R., Burton, R. A., Fincher, G. B., & Bacic, A. (2009). A barley cellulose synthase-like CSLH gene in transgenic Arabidopsis, 106(14). http://doi.org/10.1073/pnas.0902019106

Douchkov, D., Lück, S., Johrde, A., Nowara, D., Himmelbach, A., Rajaraman, J., Schweizer, P. (2014). Discovery of genes affecting resistance of barley to adapted and non-adapted powdery mildew fungi. *Genome Biology*, *15*(12), 1–18. http://doi.org/10.1186/s13059-014- 0518-8

Dwivany, F. M., Yulia, D., Burton, R.A., Shirley, N. J., Wilson, S. M., Fincher, G. B., Doblin, M.S. (2009). The CELLULOSE-SYNTHASE LIKE C (CSLC) family of barley includes members that are integral membrane proteins targeted to the plasma membrane. *Molecular Plant*, *2*(5), 1025–1039. http://doi.org/10.1093/mp/ssp064

Eckardt, N.A. (2003). Cellulose Synthesis Takes the CesA Train. *The Plant Cell*, *15*(8), 1685– 1687. http://doi.org/10.1105/tpc.150810

Ermawar, R.A., Collins, H. M., Byrt, C. S., Betts, N. S., Henderson, M., Shirley, N. J., Burton, R.A (2015). Distribution, structure and biosynthetic gene families of (1,3;1,4)-β-glucan in Sorghum bicolor. *Journal of Integrative Plant Biology*, *57*(4), 429–445. http://doi.org/10.1111/jipb.12338

Falster, D. S., & Westoby, M. (2003). Plant height and evolutionary games. *Trends in Ecology and Evolution*, *18*(7), 337–343. http://doi.org/10.1016/S0169-5347(03)00061-2

Farrokhi, N., Burton, R.A., Brownfield, L., Hrmova, M., Wilson, S. M., Bacic, A., & Fincher, G. B. (2006). Plant cell wall biosynthesis: Genetic, biochemical and functional genomics approaches to the identification of key genes. *Plant Biotechnology Journal*, *4*(2), 145–167. http://doi.org/10.1111/j.1467-7652.2005.00169.x

Fincher, G. B. (2009). Exploring the evolution of (1,3;1,4)-β-D-glucans in plant cell walls: comparative genomics can help! *Current Opinion in Plant Biology*, *12*(2), 140–7. http://doi.org/10.1016/j.pbi.2009.01.002

Fry, S. C., Nesselrode, B. H. W. a, Miller, J. G., & Mewburn, B. R. (2008). Mixed-linkage (1,3;1,4)-D-glucan is a major hemicellulose of Equisetum (horsetail) cell walls. *New Phytologist*, *179*(1), 104–115. http://doi.org/10.1111/j.1469-8137.2008.02435.x

Gardiner, J. C., Taylor, N. G., & Turner, S. R. (2003). Control of cellulose synthase complex localization in developing xylem. *The Plant Cell*, *15*(8), 1740–1748. http://doi.org/10.1105/tpc.012815

Gibeaut, D. M. (2000). Nucleotide sugars and glycosyltransferases for synthesis of cell wall matrix polysaccharides. *Plant Physiology and Biochemistry*, *38*(1-2), 69–80. http://doi.org/10.1016/S0981-9428(00)00167-4

Gille, S., Cheng, K., Skinner, M. E., Liepman, A. H., Wilkerson, C. G., & Pauly, M. (2011). Deep sequencing of voodoo lily (*Amorphophallus konjac*): An approach to identify relevant genes involved in the synthesis of the hemicellulose glucomannan. *Planta*, *234*(3), 515–526. http://doi.org/10.1007/s00425-011-1422-z

Gola, E. M. (2014). Dichotomous branching: the plant form and integrity upon the apical meristem bifurcation. *Frontiers in Plant Science*, *5*(June), 263. http://doi.org/10.3389/fpls.2014.00263

Goubet, F., Barton, C. J., Mortimer, J. C., Yu, X., Zhang, Z., Miles, G. P., & Dupree, P. (2009). Cell wall glucomannan in Arabidopsis is synthesised by CSLA glycosyltransferases, and influences the progression of embryogenesis. *Plant Journal*, *60*(3), 527–538. http://doi.org/10.1111/j.1365-313X.2009.03977.x

Graham, L. E., Cook, M. E., & Busse, J. S. (2000). The origin of plants: body plan changes contributing to a major evolutionary radiation. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(9), 4535–40

Gribaldo, S., & Brochier-Armanet, C. (2006). The origin and evolution of Archaea: a state of the art. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *361*(1470), 1007–1022. http://doi.org/10.1098/rstb.2006.1841

Haigler, C.H., Brown, R.M, & Benziman, M. (1980). Calcofluor white ST Alters the in vivo assembly of microfibrils. *Science*, *210*(4472), 903-906. http://dx.doi.org/10.1126/science.7434003

Hamann, T., Osborne, E., Youngs, H. L., Misson, J., Nussaume, L., & Somerville, C. (2004). Global expression analysis of CESA and CSL genes in Arabidopsis. *Cellulose*, *11*(3-4), 279–286. http://doi.org/10.1023/B:CELL.0000046340.99925.57

Hayashi, T., Ogawa, K., & Mitsuishi, Y. (1994). Characterization of the adsorption of Xyloglucan to Cellulose. *Plant Cell Physiol.*, *35*(8), 1199–1205. Retrieved from http://pcp.oxfordjournals.org/content/35/8/1199.short

Hazen, S. P., Scott-Craig, J. S., & Walton, J. D. (2002). Cellulose synthase-like (CSL) genes of rice Cellulose Synthase-Like Genes of Rice 1, (January).

http://doi.org/10.1104/pp.010875.synthesize

Hsieh, Y. S., & Harris, P. J. (2009). Xyloglucans of monocotyledons have diverse structures. *Molecular Plant*, *2*(5), 943-965. http://dx.doi.org/10.1093/mp/ssp061

Hrmova, M., Farkas, V., Lahnstein, J., & Fincher, G.B. (2007). A Barley Xyloglucan Xyloglucosyl Transferase Covalently Links Xyloglucan, Cellulosic Substrates, and (1,3;1,4)-β-D- Glucans. *Journal of Biological Chemistry 282*(17), 12951-12962.

Igura, M., Maita, N., Kamishikiryo, J., Yamada, M., Obita, T., Maenaka, K., & Kohda, D. (2008). Structure-guided identification of a new catalytic motif of oligosaccharyltransferase. *The EMBO Journal*, *27*(1), 234–243. http://doi.org/10.1038/sj.emboj.7601940

Kerstens, S., Decraemer, W. F., & Verbelen, J. P. (2001). Cell walls at the plant surface behave mechanically like fiber-reinforced composite materials. *Plant Physiology*, *127*(2), 381–385. http://doi.org/10.1104/pp.010423

Kim, S. J., Zemelis, S., Keegstra, K., & Brandizzi, F. (2015). The cytoplasmic localization of the catalytic site of CSLF6 supports a channeling model for the biosynthesis of mixed-linkage

glucan. *The Plant Journal*, *81*(4), 537-547.

Köhnke, T., Östlund, Å., & Brelid, H. (2011). Adsorption of arabinoxylan on cellulosic surfaces: Influence of degree of substitution and substitution pattern on adsorption characteristics. *Biomacromolecules*, *12*(7), 2633–2641. http://doi.org/10.1021/bm200437m

Kurek, I., Kawagoe, Y., Jacob-Wilk, D., Doblin, M., & Delmer, D. (2002). Dimerization of cotton fiber cellulose synthase catalytic subunits occurs via oxidation of the zinc-binding domains. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11109–11114. http://doi.org/10.1073/pnas.162077099

Labandeira, C. C. (2006). The four phases of plant-arthropod associations in deep time. *Geologica Acta*.

Lairson, L. L., Henrissat, B., Davies, G. J., & Withers, S. G. (2008). Glycosyltransferases: structures, functions, and mechanisms. *Annual Review of Biochemistry*, *77*, 521–55. http://doi.org/10.1146/annurev.biochem.76.061005.092322

Lewis, L. a, & McCourt, R. M. (2004). Green algae and the origin of land plants. *American Journal of Botany*, *91*(10), 1535–56. http://doi.org/10.3732/ajb.91.10.1535

Liepman, A. H., & Cavalier, D. M. (2012). The CELLULOSE SYNTHASE-LIKE A and CELLULOSE SYNTHASE-LIKE C families: recent advances and future perspectives.

Frontiers in Plant Science, *3*(May), 1–7. http://doi.org/10.3389/fpls.2012.00109

Liepman, A. H., Nairn, C. J., Willats, W. G. T., Sørensen, I., Roberts, A. W., & Keegstra, K. (2007). Functional genomic analysis supports conservation of function among cellulose synthase-like a gene family members and suggests diverse roles of mannans in plants. *Plant Physiology*, *143*(4), 1881–1893. http://doi.org/10.1104/pp.106.093989

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, *42*(D1), 490–495. http://doi.org/10.1093/nar/gkt1178

Mohnen, D. (2008). Pectin structure and biosynthesis. *Current opinion in plant biology, 11*(3), 266-277. http://doi.org/10.1016/j.pbi.2008.03.006.

Morgan, J. L. W., McNamara, J. T., & Zimmer, J. (2014). Mechanism of activation of bacterial cellulose synthase by cyclic di-GMP. *Nature Structural & Molecular Biology*, *21*(5), 489–96. http://doi.org/10.1038/nsmb.2803

Morgan, J. L. W., Strumillo, J., & Zimmer, J. (2013). Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature*, *493*(7431), 181–6. http://doi.org/10.1038/nature11744

Morrall, P., & Briggs, D.E. (1978). Changes in cell wall polysaccharides of germinating barley grains. *Phytochemistry, 17*(9), 1495-1502. http://doi.org/10.1016/S0031-9422(00)94628-4

Newman, R.H., Hill, S.J., & Harris, P.J. (2013). Wide-Angle X-Ray Scattering and Solid-State Nuclear Magnetic Resonance Data Combined to Test Models for Cellulose Microfibrils in Mung Bean Cell Walls. *Plant Physiology*, *163*(4), 1558–1567. http://dx.doi.org/10.1104/pp.113.228262

Niklas, K. J. (2004). Computer Models of Early Land Plant Evolution. *Annual Review of Earth and Planetary Sciences*, *32*(1), 47–66. http://doi.org/10.1146/annurev.earth.32.092203.122440

Niklas, K.J. (2004). The Cell Walls that Bind the Tree of Life. *BioScience*. http://doi.org/10.1641/0006-3568(2004)054[0831:TCWTBT]2.0.CO;2

Niklas, K. J., & Kutschera, U. (2009). The evolutionary development of plant body plans. *Functional Plant Biology*, *36*(8), 682–695. http://doi.org/10.1071/FP09107

Nobles, D. R., & Brown, R. M. (2004). The pivotal role of cyanobacteria in the evolution of cellulose synthases and cellulose synthase-like proteins. *Cellulose*, *11*(3/4), 437–448. http://doi.org/10.1023/B:CELL.0000046339.48003.0e

Nobles, D. R., Romanovicz, D. K., & Brown, R. M. (2001). Cellulose in Cyanobacteria . Origin of Vascular Plant Cellulose Synthase?, *127*(October), 529–542. http://doi.org/10.1104/pp.010557.Decho

Oehme, D.P., Downton, M.T., Doblin, M.S., Wagner, J., Gidley, M.J., & Bacic, T. (2015). Novel aspects of the structure and dynamics of Iβ elementary cellulose microfibrils revealed by computational simulations. *Plant Physiology*, *168*(1), 3–17. http://dx.doi.org/10.1104/pp.114.254664

Palin, R., & Geitmann, A. (2012). The role of pectin in plant morphogenesis. *BioSystems*, *109*(3), 397–402. http://doi.org/10.1016/j.biosystems.2012.04.006

Park, Y. B., & Cosgrove, D. J. (2015). Xyloglucan and its Interactions with Other Components of the Growing Cell Wall. *Plant and Cell Physiology*, *56*(2), 180–194. http://doi.org/10.1093/pcp/pcu204

Pauly, M., Albersheim, P., Darvill, A., & York, W. S. (1999). Molecular domains of the cellulose/xyloglucan network in the cell walls of higher plants. *Plant Journal*, *20*(6), 629– 639. http://doi.org/10.1046/j.1365-313X.1999.00630.x

Pear, J. R., Kawagoe, Y., Schreckengost, W. E., Delmer, D. P., & Stalker, D. M.

(1996). Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(22), 12637–12642. http://doi.org/10.1073/pnas.93.22.12637

Peña, M. J., Darvill, A. G., Eberhard, S., York, W. S., & O'neill, M. A. (2008). Moss and liverwort xyloglucans contain galacturonic acid and are structurally distinct from the xyloglucans synthesized by hornworts and vascular plants. *Glycobiology*, *18*(11), 891-904. https://doi.org/10.1093/glycob/cwn078

Popper, Z.A., Michel, G., Hervé, C., Domozych, D. S., Willats, W. G. T., Tuohy, M. G., Kloareg, B., & Stengel, D. B. (2011). Evolution and diversity of plant cell walls: from algae to flowering plants. *Annual Review of Plant Biology*, *62*, 567–590. http://doi.org/10.1146/annurev-arplant-042110-103809

Popper, Z.A. (2003). Primary Cell Wall Composition of Bryophytes and Charophytes. *Annals of Botany*, *91*(1), 1–12. http://doi.org/10.1093/aob/mcg013

Richmond, T. a, & Somerville, C. R. (2000). The cellulose synthase superfamily. *Plant Physiology*, *124*(2), 495–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1539280&tool=pmcentrez&rendertype=abstract

Roberts, A. W., & Roberts, E. (2004). Cellulose synthase (CesA) genes in algae and seedless plants. *Cellulose*, *11*(3/4), 419–435. http://doi.org/10.1023/B:CELL.0000046418.01131.d3

Ryden, P., Sugimoto-Shirasu, K., Smith, A. C., Findlay, K., Reiter, W.-D., & McCann, M. C. (2003). Tensile properties of Arabidopsis cell walls depend on both a xyloglucan cross- linked microfibrillar network and rhamnogalacturonan II-borate complexes. *Plant Physiology*, *132*(2), 1033–1040. http://doi.org/10.1104/pp.103.021873

Sarkar, P., Bosneaga, E., & Auer, M. (2009). Plant cell walls throughout evolution: Towards a molecular understanding of their design principles. *Journal of Experimental Botany*, *60*(13), 3615–3635. http://doi.org/10.1093/jxb/erp245

Scheller, H. V., & Ulvskov, P. (2010). Hemicelluloses. *Annual Review of Plant Biology*, *61*, 263– 289. http://doi.org/10.1146/annurev-arplant-042809-112315

Schreiber, M., Wright, F., MacKenzie, K., Hedley, P. E., Schwerdt, J. G., Little, A., Burton, R.A., Fincher, G.B., Marshall, D., Waugh, R., & Halpin, C. (2014). The barley genome sequence assembly reveals three additional members of the CslF (1,3;1,4)-β-glucan synthase gene family. *PloS One*, *9*(3), e90888. http://doi.org/10.1371/journal.pone.0090888

Scott, A. C., & Glasspool, I. J. (2006). The diversification of Paleozoic fire systems and fluctuations in atmospheric oxygen concentration. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(29), 10861–5. http://doi.org/10.1073/pnas.0604090103

Seifert, G. J. (2004). Nucleotide sugar interconversions and cell wall biosynthesis: how to bring the inside to the outside. *Current opinion in plant biology*, *7*(3), 277-284. http://doi.org/10.1016/j.pbi.2004.03.004

Shimizu-Sato, S., & Mori, H. (2001). Control of outgrowth and dormancy in axillary buds. *Plant Physiology*, *127*(4), 1405–1413. http://doi.org/10.1104/pp.010841

Somerville, C. (2006). Cellulose synthesis in higher plants. *Annual Review of Cell and Developmental Biology*, *22*, 53–78. http://doi.org/10.1146/annurev.cellbio.22.022206.160206

Sørensen, I., Domozych, D., & Willats, W. G. T. (2010). How have plant cell walls evolved? *Plant Physiology*, *153*(2), 366–372. http://doi.org/10.1104/pp.110.154427

Sussex, I., & Kerk, N. (2001). The evolution of plant architecture. *Current Opinion in Plant Biology*, *4*(1), 33–7. http://doi.org/10.1016/S1369-5266(00)00132-1

Suzuki, S., Li, L., Sun, Y.-H., & Chiang, V. L. (2006). The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in Populus trichocarpa. *Plant Physiology*, *142*(3), 1233–1245. http://doi.org/10.1104/pp.106.086678

Tsekos, I. (1999). The sites of cellulose synthesis in algae: diversity and evolution of cellulose-synthesizing enzyme complexes. *Journal of phycology*, *35*(4), 635-655. http://doi.org/10.1046/j.1529-8817.1999.3540635.x

Verhertbruggen, Y., Yin, L., Oikawa, A., & Scheller, H. V. (2011). Mannan synthase activity in the CSLD family. *Plant Signaling & Behavior*, *6*(10), 1620–1623. http://doi.org/10.4161/psb.6.10.17989

Vincken, J.-P., Schols, H.A, Oomen, R. J. F. J., McCann, M. C., Ulvskov, P., Voragen, A. G. J., & Visser, R. G. F. (2003). If homogalacturonan were a side chain of rhamnogalacturonan I. Implications for cell wall architecture. *Plant Physiology*, *132*(4), 1781–1789. http://doi.org/10.1104/pp.103.022350

Wang, D. Y., Kumar, S., & Hedges, S. B. (1999). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proceedings. Biological Sciences / The Royal Society*, *266*(1415), 163–171. http://doi.org/10.1098/rspb.1999.0617

Whitney, S. E. C., Gothard, M. G. E., Mitchell, J. T., & Gidley, M. J. (1999). Roles of Cellulose and Xyloglucan in Determining the Mechanical Properties of Primary Plant Cell Walls1. *Plant Physiology*, *121*(2), 657–664. http://doi.org/10.1104/pp.121.2.657

Willats, W. G. T., Orfila, C., Limberg, G., Buchholt, H. C., Van Alebeek, G. J. W. M., Voragen, A.G. J., … Knox, J. P. (2001). Modulation of the degree and pattern of methyl-esterification of pectic homogalacturonan in plant cell walls: Implications for pectin methyl esterase action, matrix properties, and cell adhesion. *Journal of Biological Chemistry*, *276*(22), 19404– 19413. http://doi.org/10.1074/jbc.M011242200

Wilson, S. M., Ho, Y. Y., Lampugnani, E. R., Van de Meene, A. M. L., Bain, M. P., Bacic, A., & Doblin, M. S. (2015). Determining the Subcellular Location of Synthesis and Assembly of the Cell Wall Polysaccharide (1,3; 1,4)-β-D-Glucan in Grasses. *The Plant Cell Online*, *27*(March), tpc.114.135970. http://doi.org/10.1105/tpc.114.135970

Yin, Y., Huang, J., & Xu, Y. (2009). The cellulose synthase superfamily in fully sequenced plants and algae. *BMC Plant Biology*, *9*, 99. http://doi.org/10.1186/1471-2229-9-99

Zykwinska, A. W., Ralet, M.-C. J., Garnier, C. D., & Thibault, J.-F. J. (2005). Evidence for *in vitro* binding of pectin side chains to cellulose. *Plant Physiology*, *139*(1), 397–407. http://doi.org/10.1104/pp.105.065912

Statement of Authorship

Title of Paper	Evolutionary Dynamics of the Cellulose Synthase Gene Superfamily in Grasses	
Publication Status	☑ Published □ Submitted for Publication	□ Accepted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Published in Plant Physiology: ORCID IDs: 0000-0002-9567-1856 (K.M.); 0000-0002-3488-0531 (A.J.H.); 0000-0002-1808-8130 (C.H.)	

Principal Author

Name of Principal Author (Candidate)	Julian Schwerdt		
Contribution to the Paper	Acquired and curated sequence data. Performed phylogenetic and selection analyses. Performed homology modelling.		
Overall percentage (%)	60		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	27/5/16

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Frank Wright		
Contribution to the Paper	Contributed to phylogenetic and selection analyses.		
Signature		Date	27/6/2016

Name of Co-Author	Katrin MacKenzie		
Contribution to the Paper	Contributed to phylogenetic and selection analyses.		
Signature		Date	27.5.16

Name of Co-Author	Daniel Oehme		
Contribution to the Paper	Performed molecular dynamics simulations.		
Signature		Date	27/05/2016

Name of Co-Author	John M. Wagner		
Contribution to the Paper	Performed molecular dynamics simulations.		
Signature		Date	27 May 2016

Name of Co-Author	Andrew J. Harvey		
Contribution to the Paper	Identified conserved insert in *CslF6.*		
Signature		Date	06/06/2016

Name of Co-Author	Neil J. Shirley		
Contribution to the Paper	Defined chromosomal locations of genes.		
Signature		Date	2/8/2016

Name of Co-Author	Rachel A. Burton		
Contribution to the Paper	Defined chromosomal locations of genes.		
Signature		Date	2/8/16

Name of Co-Author	Miriam Schreiber		
Contribution to the Paper	Defined chromosomal locations of genes.		
Signature		Date	30/05/2016

Name of Co-Author	Claire Halpin		
Contribution to the Paper	Defined chromosomal locations of genes.		
Signature		Date	27 - 5- 2016

Name of Co-Author	Jochen Zimmer		
Contribution to the Paper	Provided atomic coordinates and assisted with structural biology.		
Signature		Date	5/27/16

Name of Co-Author	David F. Marshall		
Contribution to the Paper	Defined chromosomal locations of genes. Contributed to phylogenetic analyses. Conceived project and designed project and experiments.		
Signature		Date	25 - 6- 2016

Name of Co-Author	Robbie Waugh		
Contribution to the Paper	Conceived project and designed project and experiments.		
Signature		Date	14.7.2016

Name of Co-Author	Geoffrey B. Fincher		
Contribution to the Paper	Conceived project and designed project and experiments.		
Signature		Date	2/8/16.

Schwerdt, J.G., MacKenzie, K., Wright, F., Oehme, D., Wagner, J.M., Harvey, A.J., Shirley, N.J., Burton, R.A., Schreiber, M., Halpin, C., Zimmer, J., Marshall, D.F., Waugh, R. & Fincher, G.B. (2015). Evolutionary dynamics of the cellulose synthase gene superfamily in grasses.
Plant Physiology, *168*(3), 968-983.

Statement of Authorship

Title of Paper	Cellulose synthase-like J (CslJ) genes constitute a third lineage to encode (1,3;1,4)-β-glucan synthases in angiosperms
Publication Status	□ Published □ Accepted for Publication □ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Julian Schwerdt
Contribution to the Paper	Performed all phylogenetic, selection and genomic analyses. Prepared manuscript.
Overall percentage (%)	40
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 2/8/16

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Alan Little
Contribution to the Paper	Heterologous expression of Nicotiana Benthamiana. Assisted with manuscript.
Signature	Date 2/8/16

Name of Co-Author	Jelle Lahnstein
Contribution to the Paper	Performed all HPLC work.
Signature	Date 2/8/16

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Lisa O'Donovan		
Contribution to the Paper	Performed all microscopy work.		
Signature		Date	2/8/2016.

Name of Co-Author	Neil Shirley		
Contribution to the Paper	Performed all QPCR work.		
Signature		Date	2/8/2016

Name of Co-Author	Geoffrey B. Fincher		
Contribution to the Paper	Conceived project and designed experiments.		
Signature		Date	2/8/16.

Name of Co-Author	Hellen M. Collins		
Contribution to the Paper	Performed all transgenic overexpression work.		
Signature		Date	2/8/2016

Name of Co-Author	Rachel A. Burton		
Contribution to the Paper	Conceived project and designed experiments. Assisted with manuscript.		
Signature		Date	2/8/16

Running Title:

The barley cellulose synthase-like *CslJ* gene encodes a (1,3;1,4)-β-glucan synthase

Corresponding author: Rachel Burton

ARC Centre of Excellence in Plant Cell Walls, School of Agriculture, Food and Wine,
University of Adelaide, Waite Campus, Glen Osmond, SA 5064, Australia

Phone: +61 8 8313 1057

Fax: +61 8313 7102

Email: rachel.burton@adelaide.edu.au

Cellulose synthase-like J (*CslJ*) genes constitute a third lineage to encode (1,3;1,4)-β-glucan synthases in angiosperms

Julian Schwerdt[1], Alan Little[1], Jelle Lahnstein[2], Lisa O'Donovan[2], Neil Shirley[2], Geoffrey B. Fincher[2], Helen M. Collins[2] and Rachel A. Burton[3].

[1,2,3]ARC Centre of Excellence in Plant Cell Walls, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Glen Osmond, SA 5064, Australia.

[1]These authors contributed jointly to the work

[3]Corresponding author

Key Words:

(1,3;1,4)-β-glucan biosynthesis, *cellulose synthase-like genes,* convergent evolution, selection pressure

ABSTRACT

In plants the presence of the cell wall polysaccharide (1,3;1,4)-β-glucan is distributed across embryophytes but is predominantly found in the Poaceae, which include many economically important cereal and grass species. Of the three Poales-specific cellulose synthase-like gene families, *CslF* and *CslH* have been shown previously to encode (1,3;1,4)-β-glucan synthases. Here, we demonstrate that *HvCslJ1* also mediates (1,3;1,4)-β-glucan synthesis when transiently expressed in *Nicotiana benthamiana*, although no changes in (1,3;1,4)-β-glucan content were observed in grain from stably transformed plants overexpressing *HvCslJ1.* This suggested that an additional factor essential for (1,3;1,4)-β-glucan synthesis was not present in the endosperm of the transgenic grain. Transcript profiling indicated that endogenous *HvCslJ1* mRNA is more abundant in vegetative tissues, particularly roots and embryos, than in developing or mature grain endosperm. Phylogenetic analyses recovered *CslJ* genes from the Poaceae as strongly monophyletic and revealed highly divergent evolutionary histories between them and a large monophyletic lineage of eudicot *CslJ* relatives. We propose to name this monophyletic eudicot lineage cellulose synthase-like M (*CslM*). The identification of a third cellulose synthase-like gene family capable of (1,3;1,4)-β-glucan biosynthesis confirms that the evolution of (1,3;1,4)-β-glucan synthases has occurred on multiple independent occasions in land plants and at least three times in the grasses, and that it provides a valuable case study in convergent evolution.

INTRODUCTION

Knowledge of the evolutionary history of the cellulose synthase gene superfamily has grown with the increased availability of fully sequenced plant genomes [1, 2]. However, determining the functional role of each lineage in the synthesis of specific plant cell wall polysaccharides has been hindered by the complex biochemical nature of cell wall biosynthesis and the difficulty posed by working with membrane-bound proteins during *in vitro* functional characterization experiments. The gene superfamily comprises ten nominal groupings, namely *CesA, CslA, CslB, CslC, CslD, CslE, CslF, CslG, CslH* and *CslJ*. Functional evidence linking the synthesis of cellulose to *CesA* genes*,* the synthesis of mannan to *CslA* and the synthesis of xyloglucan to *CslC* [3–5] is consistent with the position of these polysaccharides on the archaeplastida tree and with the wide taxonomic representation of those clades [6]. Cellulose, xyloglucan and mannan appear as retained ancestral traits, although mannan synthesis was lost in many embryophyte lineages following the split between the spermatophytes and bryophytes [7]. In contrast, (1,3;1,4)-β-glucan, a polymer of (1,4)-β-glucosyl residues containing single, interspersed (1,3)-β-linked glucosyl residues [8], appears in just a few eukaryotes, where it is distributed widely in grasses (Poaceae), but more restricted in the horsetail fern *Equisetum sp.,* the fungus *Rhynchosporium secalis,* the red algae *Kappaphycus* and some brown algae, among others [9–13]. These observations have prompted the suggestion, based on the parsimony principle, that (1,3;1,4)-β-glucan synthesis has multiple independent origins [6,14].

As noted above, (1,3;1,4)-β-glucan is primarily observed in cell walls of the Poaceae and is likely to be present in the majority of grass species [15]. Previous taxonomic sampling has identified three Poaceae-specific cellulose synthase superfamily lineages, namely *CslF, CslH* and *CslJ*, and these are therefore natural candidate genes for (1,3;1,4)-β-glucan synthases [16]. The recent release of the pineapple genome [17] has revealed that *CslH* and *CslJ* genes occur in the bromeliads but no *CslF* genes were reported. This shows the *CslF* family to be the only grass-specific family whilst the other families are observed in the Poales, which include the grasses and the bromeliads. The tight association of (1,3;1,4)-β-glucan to Poales in the spermatophytes provides the opportunity to express potential (1,3;1,4)-β-glucan synthases in an embryophyte heterologous system devoid of (1,3;1,4)-β-glucan. This is essential in unambiguously defining the polysaccharide products [16].

To date, both the *HvCslF* and *HvCslH* family genes have been shown to encode enzymes that synthesise (1,3;1,4)-β-glucan in both monocot and heterologous eudicot systems [8,16]. The most widely tested and productive so far is *HvCslF6*, which produced over 20-fold higher levels of (1,3;1,4)-β-glucan when expressed in transient and transgenic systems, compared with other *CslF* or *CslH* family members [8]. Although Poaceae-specific, *CslF* and Poales specific *CslH* are from deeply diverged lineages, at least as old as the ~140-150 Mya (million years ago) monocot-eudicot split [18]. When considered alongside the scattered distribution of (1,3;1,4)-β-glucan across archaeplastida, (1,3;1,4)-β-glucan synthesis in plants appears likely to have emerged independently through convergent evolution.

In an effort to explore the origin of (1,3;1,4)-β-glucan synthesis and to expand the genetic tools available for *in planta* modification of (1,3;1,4)-β-glucan structure and abundance, we have tested the remaining Poales-specific family, *CslJ*, for (1,3;1,4)-β-glucan synthase activity [19] and demonstrate that it has equivalent synthase activity compared with *HvCslH* in a heterologous transient expression system. However, the status of *CslJ* as a Poales-specific gene family has been challenged by phylogenetic analyses of eudicot genes that cluster wit*h CslJ* and are distinct from *CslG* [1]. Thus, Yin et al. [1] concluded that CslJ genes were found in both monocots and dicots*.* In an attempt to resolve the systematic status of *CslJ* genes we have conducted phylogenetic and genomic analyses of thirteen eudicot and eight monocot species. Additionally, we have compared the selection forces operating on *CslJ* genes and their closest eudicot relatives since the appearance of their most recent common ancestor.

Our phylogenetic analyses reveal a large monophyletic lineage of eudicot *CslJ*-related genes that appear to have undergone a pattern of gene duplication similar to both *CslG* and the more distant *CslE* genes*.* However, the *CslJ* clade of the Poaceae is conspicuous by its comparatively small number of duplication events, conserved gene structure and significant shifts in selection pressure following the eudicot-monocot divergence. These results support a functional and nomenclatural division of *CslJ genes* from their closest eudicot relatives and *CslG* clades.

RESULTS

Transient expression of *HvCslJ1* in *Nicotiana benthamiana* results in detectable levels of (1,3;1,4)-β-glucan

The *HvCslJ1,* a *CslG* gene from *Vitis vinifera* (grapevine; *VvCslG*; VIT_05s0020g05050) and the positive control genes *HvCslF6 and HvCslH1 from barley*, all driven by the *35S* promoter, were introduced into *Nicotiana benthamiana* leaves using *Agrobacterium* infiltration. The binary expression vector also carried the p19 silencing suppressor in order to maximise expression without any degradation of transcript due to RNA silencing [20]. Six days after infiltration samples were collected for detection of (1,3;1,4)-β-glucan using biochemical assays and transmission electron microscopy. Infiltrated leaf samples were digested with an endohydrolase that specifically targets (1,3;1,4)-β-glucan, through the hydrolysis of (1,4)-β-linkages immediately towards the reducing terminus from (1,3)-β-linkages, and that releases characteristic oligosaccharides. The high performance anion-exchange chromatography (HPAEC) profile for *HvCslJ1*-infiltrated tissues contained diagnostic DP3 and DP4 oligosaccharides, where DP denotes degree of polymerization, as seen in the standard control (Figure 2A), indicating that *HvCslJ1* was capable of producing (1,3;1,4)-β-glucan (Figure 2D). The amount of (1,3;1,4)-β-glucan measured by enzymic digestion was relatively low at a maximum of 0.1% as compared with 1.6% produced by *HvCslF6* (Figure 2B), and was similar to the amount generated by *HvCslH1* at a maximum of 0.05% (Figure 2C). There were minor variations in the DP3:DP4 ratios of the (1,3;1,4)-β-glucan synthesised by the three different genes with *HvCslF6* at 1.6:1 (Figure 2B), *HvCslH1* at 1.4:1 (Figure 2C) and *HvCslJ1* at 1:3.1 (Figure 2D).

The presence of (1,3;1,4)-β-glucan in the *N. benthamiana* leaf following expression of *HvCslJ1* was confirmed by probing the fixed and embedded infiltrated tissue with the (1,3;1,4)-β-glucan-specific antibody BG1, followed by transmission electron microscopy detection using a secondary antibody conjugated to 18nm gold particles. Labelling was detected in the walls of the leaf tissue infiltrated with *HvCslJ1* at a level consistent with the HPAEC quantification (Figure 3B). No labelling was observed in *N. benthamiana* leaves infiltrated with wild type AGL1 (Figure 3A).

CslJ family members form a highly diverged clade nested within large eudicot expansions

Bayesian phylogenetic analysis (Figure 5A) of the *CslE, CslG* and *CslJ* gene families from thirteen eudicot and eight monocot species reveals four major lineages. The *CslE* family is comprised of a basal divergence between a small clade of eudicot representatives and a larger clade containing reciprocally monophyletic monocot and eudicot groupings. Paralogous clusters are found specific to species or closely related species-groups throughout the family, for instance after the Pooideae split from other Poaceae. The *CslG* family forms a clade with multiple duplication events within all sampled higher plant taxa. Our analyses further detected *Musa acuminata* (banana) and *Panicum virgatum* (switchgrass) genes within the *CslG* clade (as sister to all other sampled lineages), in contrast to expectations that this *CslG* gene family is eudicot-specific [21]. That the node uniting *CslG* and *CslJ* is poorly supported (0.49 posterior probability) in the Bayesian tree might indicate a problematic placement of these monocot genes in *CslG*. However, this node is well supported in our maximum likelihood analysis (Supplementary Figure S1) and also in a DensiTree (Supplementary Figure S2) plot of trees sampled in the Bayesian analysis to visualise the distribution of alternative sampled topologies. This shows that the placement of the *CslG* family accounts for the conflicting topologies but that the banana and switchgrass genes remain clustered with this clade.

The Poaceae *CslJ* genes comprise a clade that is separated by a long molecular branch from its closest sampled *CslJ* relatives; the latter form a large eudicot-specific clade defined in this study as a new cellulose synthase-like family called M (*CslM)* (Figure 5). Notung [22] reconciliation of species and gene trees for the *CslJ* family and the newly defined *CslM* genes showed contrasting histories of gene duplications and losses in these clades*.* The eudicot-specific *CslM* genes underwent multiple duplication events early in the history of this group and subsequently within several sampled species, and experienced frequent gene losses in each major subclade (Figure 5B). In contrast, the Poaceae *CslJ* family appears to have undergone few losses, primarily in the rice and *Brachypodium* lineages, one basal gene duplication after the Panicoideae divergence and a paralogous duplication in *Setaria italica.* As shown in Figure 6, an analysis of *CslJ* and *CslM* gene structure is consistent with comparatively more genomic reorganisation in *CslM.* A GenePainter [23] plot mapped onto a major lineage tree reveals significant intron gain and loss throughout the *CslM* family's history, in contrast to *CslJ* genes, which have had one intron loss and one gain (Figure 6A).

The *CslJ* clade has experienced a sustained shift in selection pressure following the eudicot-monocot divergence

To test how selection has operated on the Poaceae *CslJ* family following the eudicot-monocot divergence we estimated ratios of non-synonymous and synonymous substitutions (dN:dS or ω) using the branch-site model of codeml [24]. As indicated in Figure 5B, six major basal clades (A-F) were tested, with positive selection detected only in the *CslJ* lineage (F). *CslJ* is shown to have undergone a sustained shift in selective pressure across 12 amino acid sites (BEB, posterior probability >0.95) since the ancestral duplication that separated *CslJ* from *CslM* (Supplementary Table 2). Figure 7 shows that these residues are distributed across a conserved region starting from the DxD to just after the QxxRW catalytic motifs.

CslJ is transcribed in barley and sorghum roots and late in developing grain.

The transcript levels of *CslJ* were quantified across various barley and sorghum vegetative and grain tissues. A low level of transcription of *HvCslJ1* was detected in all barley vegetative tissues tested by QPCR (Figure 1A), whilst *SbCslJ1* transcript levels were markedly higher in root tissues (Figure 1D). Transcript levels were extremely low in the isolated endosperm tissues from barley developing grain from 6 to 38 days after pollination (DAP) (Figure 1B) but were higher in whole developing grain samples, where all tissues were included, of both barley (Figure 1B) and sorghum (Figure 1E). Analysis of RNAseq data for sorghum also confirmed the presence of much higher *SbCslJ1* transcript levels specifically in the embryonic tissues of the developing grain (Figure 1F). Additional transcript profiles were obtained from RNAseq data publically available from the sequence read archive (SRA) [25] which confirmed the highest *HvCslJ1* transcript levels were in roots and the embryo of germinated grain (Figure 1C). At such an early stage of grain germination, the embryo would be expected to contain numerous tissues including coleorhiza, coleoptile, epiblast, mesocotyl, embryonic axis and coleoptile.

HvCslJ1 and *HvCslH1* overexpression in transgenic barley grain

Multiple independent stable transgenic barley lines overexpressing *HvCslJ1* or *HvCslH1* under the control of either an endosperm-specific oat globulin promoter, *As*GLO (Vickers et al., 2003) or the constitutive promotor 35S, were generated (Supplementary Table S1). A subset of lines was selected and the T_2 grain were tested for (1,3;1,4)-β-glucan content (Figure 4A and B) and DP3:DP4 ratio (Figure 4D). Expression of *HvCslJ1* in transgenic barley grain driven by either *As*GLO or 35S resulted in no significant increases in (1,3;1,4)-β-glucan amount or changes in DP3:DP4 ratios (Figure 4A, B and D). The *As*GLO:*HvCslH1* plants produced modest increases in grain (1,3;1,4)-β-glucan amounts at an average of 5.3% (range 3.9-7.5%) compared with the wild type (4.9%) (Figure 4A), but at a significantly lower level than previously reported in *As*GLO:*HvCslF6* plants, which produced grain containing up to 7.8% (1,3;1,4)-β-glucan [26] with a concurrent decrease in

DP3:DP4 ratio down to 2.1%±0.11 [26]. There was also a small change in the DP3:DP4 ratio in *As*GLO:*HvCslH1* grain (Figure 4D). The average grain size of three of the *As*GLO:*HvCslH1* lines was significantly lower than the wild type (Figure 4C), but this phenotype was not present in any *HvCslJ1* lines.

DISCUSSION

Our study uses phylogenetic and expression analyses to shed light on the evolutionary history and functional characterization of (1,3;1,4)-β-glucan. (1,3;1,4)-β-Glucan is unique among the major cell wall polysaccharides in its highly asymmetric distribution among archaeplastida members. In embryophytes, it is primarily represented in the Poaceae with the only other significant observation in the highly divergent pteridophyte, *Equisetum*, which shows clear differences in (1,3;1,4)-β-glucan structural composition [27]. Thus, multiple independent origins of (1,3;1,4)-β-glucan synthesis is the most parsimonious scenario. Additionally, that two Poales specific cellulose synthase superfamily clades, *CslF* and *CslH,* have been shown to mediate (1,3;1,4)-β-glucan synthesis [8,16] is notable considering their divergence is at least as old as the ~140-150 Mya monocot-eudicot division [18]. This again suggests multiple independent origins of (1,3;1,4)-β-glucan synthase genes, not only across the embryophytes, but within the Poales and presents a promising potential case study in convergent evolution. The *CslJ* clade was the third *cellulose synthase-like* family that has been recognised as Poales specific [19], although this has been challenged with wider taxonomic sampling [1]. While closer to *CslH* than *CslF*, *CslJ* is nevertheless a member of a separate lineage whose origin lies before the monocot-eudicot split [1,2]. As shown in figures 2 and 3, the barley *HvCslJ1* gene encodes a protein that is capable of synthesising

(1,3;1,4)-β-glucan in the heterologous host *N. benthamiana*. As with the *Hv*CSLF6 and *Hv*CSLH1 proteins [8,16], *Hv*CSLJ1 is predicted to function independently, and although we cannot rule out the possibility that an ancillary protein or cofactor present in the *N. benthamiana* leaves may be required for activity, these transient heterologous expression results indicate that *Hv*CSLJ1 is at least as active as *Hv*CSLH1 in (1,3;1,4)-β-glucan synthesis. Thus, (1,3;1,4)-β-glucan synthesis has now been demonstrated to be mediated by genes in each of the three main cellulose synthase superfamily lineages*, namely CesA/CslD/CslF, CslB/CslH and CslE/CslG/CslJ/CslM* by *CslF, CslH* an*d CslJ*, respectively. It should be noted that the *CslA and CslC* clades are thought to have evolved from a separate endosymbiotic transfer and jointly constitute an independent lineage [28].

The manipulation or mutation of the *CslF6* gene to either increase or decrease levels of (1,3;1,4)-β-glucan has now been reported several times and it is clear that this gene encodes the dominant (1,3;1,4)-β-glucan synthase in plants [26,29–32] Here, stable transgenic barley lines were generated carrying *HvCslH1* or *HvCslJ1* genes driven by either the strong 35S promoter or the endosperm-specific *As*GLO promoter, and (1,3;1,4)-β-glucan amounts and fine structure were analysed to allow a comparison with previously published data from *AsGLO:HvCslF6* plants [26]. Of these four sets of plants, only those transformed with the *As*GLO:*CslH1* construct showed significant increases in grain (1,3;1,4)-β-glucan amount and fine structure, although these changes were modest. This result is in contrast to the activities indicated by the heterologous expression experiments in the *N. benthamiana* leaves, where expression of *35S:HvCslJ1* generated (1,3;1,4)-β-glucan at a similar level to *35S:HvCslH1*.

A clue to this contrasting behaviour may be provided by the spatial distribution of the *CslJ* transcript. Both QPCR and RNAseq data indicate that the highest levels of

CslJ transcript within both barley and sorghum grain are found at the later stages of development around 28 days after pollination in barley or at the maturing stage for sorghum (Figure 1B and 1E), but almost none of these transcripts appear to be in the endosperm tissue (Figure 1B). Rather, the *CslJ* transcript is concentrated in the embryo in both species (Figure 1C and 1F). Pre-formed tissues in the embryo include the coleoptile and the coleorhizae and fine dissection of these organs would be required to establish if *CslJ* transcript is present in both. However, it is clear that there is strong *CslJ* expression in both barley and sorghum roots later in plant development (Figure 1C and 1D). This distribution pattern is unlike that found for either *CslF6* and *CslH1* in barley [8,33] and sorghum [34]. Although transcripts of *CslH1* have been reported both in the older leaf tissues of barley [8] and in seedling tissues of *Brachypodium distachyon* [35], they are also present in the starchy endosperm, albeit at much lower levels than *CslF6,* which is heavily involved in the synthesis of the cell walls in this tissue and many others [35].

This spatial distribution implies that *CslJ* may play a key role in (1,3;1,4)-β-glucan synthesis in non-grain tissues. However, given the absence of *CslJ* sequences in rice and *Brachypodium*, the limited information available on the structural variation of (1,3;1,4)-β-glucan across the plant body and the inherent redundancy of overlapping expression patterns, it will be difficult to unravel the drivers or selection advantages for plants in which multiple gene families have independently converged on synthesising the same polysaccharide. It could depend on the requirements for different DP3:DP4 ratios, locations, solubilities and physicochemical properties of the (1,3;1,4)-β-glucan in each tissue and at various growth stages. Mutant and knockout *CslH1* and *CslJ1* lines with an associated visible or biochemical phenotype would be extremely helpful in this regard.

Other possible reasons for the results obtained for the transgenic lines include the need for an ancillary protein or unusual co-factor for (1,3;1,4)-β-glucan synthase activity. If such an ancillary protein or co-factor were present in the *N. benthamiana* leaf cells but not in the grain, this might explain the differences in the amount of (1,3;1,4)-β-glucan synthesized in each tissue. Perhaps more likely is that the promoters used for transgenesis were not active in the embryo; the *AsGLO* promoter has a high level of specificity for the starchy endosperm [26] and the *35S* promoter is not always constitutive, particularly in monocots.

Our phylogenetic analyses locate single *CslJ* representatives in *Hordeum vulgare* (barley), *Sorghum bicolor*, *Zea mays* (maize), and *Panicum virgatum* (switchgrass), while *Setaria italica* contains a paralogous *CslJ* pair (Ermawar et al., 2015). *CslJ* members were not identified in *Oryza sativa* (rice), *Brachypodium distachyon*, or in the non-Poaceae monocots sampled, *Musa acuminate* (banana) and *Phoenix dactylifera* (date palm) but a single representative was found in *Ananas comosus* (pineapple) [17]. A large eudicot-specific clade is robustly resolved as a sister to the Poaceae *CslJ* family. Previous to our study, this eudicot clade was represented only by three *Vitis vinifera* (grape) and two *Populus trichocarpa* (poplar) sequences that were previously characterised as *CslJ* [1], challenging the Poaceae-specific status of the *CslJ* gene family. Here, our expanded taxonomic sampling reveals a long stem branch leading to the *CslJ* clade, indicating relatively high rates of substitution prior to the divergence of sampled grasses. Consistent with their deep molecular split and demonstrated functional divergence, the monocot and eudicot clades also show contrasting evolutionary histories. The *CslJ* topology is mostly concordant with the species tree: a single ancestral gene duplication is inferred following the Panicoideae divergence. One of these duplicates is subsequently lost in all but

Setaria italica and additional losses are inferred in the rice and *Brachypodium* lineages (Figure 5B).

In comparison, the eudicot clade has undergone numerous ancestral gene duplications, paralogous duplication in several taxa and substantial gene losses throughout the history of the clade (Figure 5A). Intron structure is also consistent with significantly different evolutionary histories in the two clades (Figure 6B). The *CslJ* family is shown to have lost a single intron in the Paniceae (*Panicum virgatum*) and gained a single intron in Cenchrinae (*Setaria italica*) lineages (Figure 6B). The eudicot family has experienced numerous intron gains and losses throughout its history in contrast to the relatively conservative evolution of gene structure in the *CslJ* clade. What can we infer from the contrasting evolutionary dynamics of *the Poaceae CslJ genes* and their large eudicot sister clade? The many observed gene duplications could indicate selection for increasing gene dosage, suggesting a role in stress-response or metabolic pathways. Indeed, a history of shifting environmental selection pressures could have driven the dynamics observed in the *CslJ* eudicot relatives by varying selection pressure on duplicates to incur cycles of gene losses and duplications [36]. Alternative explanations for our findings include the specialisation of existing multifunctional gene products or the fixation of polymorphisms [37,38]. However, major functional diversification events are unlikely to explain the majority of gene duplications in the *CslJ* eudicot relatives because these duplications occur within extant species. Neofunctionalisation of deep ancestral duplications is more plausible because these are accompanied by relatively few gene losses more suggestive of ancient functional divisions. However, further taxonomic sampling is needed to assess whether their systematic structure is consistent with neofunctionalisation of gene duplicates.Our study highlights a significant long term shift in selection pressure on twelve amino acid sites across all branches in *CslJ* following the eudicot-monocot split. These amino acid residues are distributed across conserved regions of the PF03552 cellulose synthase PFAM domain with some positioned close to the characteristic QxxRW motif and conserved DxD residues. These amino acid residues could potentially be directly or indirectly involved in the catalysis of (1,3;1,4)-β-glucan synthesis. However, because *CslJ* and their eudicot relatives are resolved in this study to be reciprocally monophyletic (concordant with the species tree), neofunctionalisation of a gene duplicate does not appear as the origin of (1,3;1,4)-β-glucan synthesis in *CslJ*.

These phylogenetic findings and the demonstrated synthesis by the *CslJ* family of (1,3;1,4)-β-glucan, a polysaccharide not found in eudicots, argue for a discrete nomenclature for *CslJ* genes and their eudicot relatives. We propose the name *CslM* for the larger sister eudicot clade to *the Poaceae CslJ genes. The CslJ/CslM* genes resemble the *CslH/CslB* families in their reciprocally monophyletic eudicot and monocot clades *(Figure 5A).*

All three currently recognised Poales-specific gene families have now been shown to mediate the synthesis of (1,3;1,4)-β-glucan. The most parsimonious explanation for the presence of three (1,3;1,4)-β-glucan synthases in separate Poaceae specific clades is that they have independently converged on their functions as (1,3;1,4)-β-glucan synthases. An alternative scenario involves multiple independent losses of (1,3;1,4)-β-glucan synthase function throughout the cellulose synthase superfamily tree [39]. The potential of each family to be used for increasing the amount of grain (1,3;1,4)-β-glucan and improving the solubility for human health applications has been determined [26,40], with *HvCslF6* proving to be the most effective enzyme in barley. In light of this, our identification of another *Csl* gene family capable of the biosynthesis of (1,3;1,4)-β-glucan may not only advance opportunities transgenic modification of grain dietary fibre, but also contribute to the understanding of its biosynthetic mechanism through comparative genomics. Finally, the identification of *CslM* as a eudicot specific lineage that has had a dramatically different evolutionary history to its closest relative *CslJ* is a useful candidate for future work focusing on the origin of plant cell wall polysaccharides. The *CslJ* and *CslM* clades are very closely related (Figure 5A) but their gene products clearly synthesise quite different cell wall polysaccharides.

Materials and Methods

Real-time quantitative PCR and RNAseq data analysis

The barley and sorghum tissue-specific cDNA samples used for transcript profiling were previously described in Burton et al. [41] and Ermawar et al. [34], respectively. For the collection of whole grain tissues, barley plants (cv. Sloop) were grown under standard glasshouse conditions as described in Burton et al. [41]. From these, whole grain cDNAs used for QPCR were prepared as described by Burton et al. [41]. Multiple grains on at least three different spikes were collected and combined before RNA extraction. QPCR data from these samples was combined with those of the endosperm set of cDNAs described in Burton et al. [33] and normalised as described in Burton et al. [41].

HvCslJ1 and SbCslJ1 transcript profiling was carried out using QPCR as described previously [42] using the following gene-specific primers; *HvCslJ1*-qFor GAGGAGGTCGGCTTCTTGTA, *HvCslJ1*-qRev CCATGAAGGCGTAGTAGGC, *SbCslJ1*-qFor TGGAAGATGACGTTCCAATTC, *SbCslJ1*-qRev GCTGGTGGTGGACTTTACTCA. The data were normalized against the geometric mean of the four housekeeping genes, namely glyceraldehyde 3-phosphate dehydrogenase (GAPDH), cyclophilin, tubulin and HSP70 [43].

Comprehensive tissue-specific RNAseq data sets were collected for barley and sorghum [44] from the Sequence Read Archive at the DNA Data Bank of Japan. Read assembly was performed against the genome for barley (Consortium, T. I. B. G. S., 2012) and sorghum (www.gramene.org) in CLC Genomics Workbench (QIAGEN, Aarhus, Denmark).

Vector Construction

For plant transformation the constructs containing the cDNA of *HvCslF6* in the modified pMDC32 vector pRB474 driven by either the endosperm-specific oat globulin promoter (*As*GLO) or the pMDC32 vector carrying the *35S* promotor [45] were as described in Burton et al. [26]. Total RNA extracted from various barley tissues as described previously [46] was used to synthesise cDNA as described in Burton et al. [33] and used for the amplification of cDNA fragments covering the open reading frame of both *HvCslH1* and *HvCslJ1* using the following gene-specific primers;

HvCslH1-cF1 TCGAGCGGTTGTTGCTTGTG,

HvCslH1-cR1 CCTGCTTGAGTCTTCGTTACATGTTC,

HvCslJ1-For ATGCTGGCGGCCGACCTGGCG,

HvCslJ1-Rev TTAACCAAACAAGCAAAGCAG.

The fragments were inserted into the Gateway entry vector pCR8 according to the manufacturer’s instructions and transferred to applicable destination vectors using the LR clonase reaction (Life Technologies, Carlsbad, U.S.A.). All constructs were checked by sequencing (Australian Genome Research Facility, Adelaide) prior to transformation. For *Nicotiana benthamiana* infiltration the cDNAs were transferred from the entry vector to the pEAQ-HT-DEST1 destination vector [47]. Plasmids were used to transform the *Agrobacterium tumefaciens* strain AGL1 by a freeze-thaw method [48].

Transient heterologous expression in *Nicotiana benthamiana*

All transient heterologous expression experiments in *N. benthamiana* plants were carried out as detailed in [40].

Agrobacterium tumefaciens-mediated barley transformation

The barley cultivar 'Golden Promise' was transformed by infection with *Agrobacterium* cells carrying one of the *HvCslH1* and *HvCslJ1* constructs and the transgenic lines, through to the T2 generation, were confirmed at the genotype level as described in Burton et al. [26] (data not shown). A total of 17, 19, 20 and 19 lines were successfully regenerated carrying the *As*GLO*::HvCslH1, 35S::HvCslH1, As*GLO*::HvCslJ1* and *35S::HvCslJ1,* respectively (Supplementary Table S1). Of these, six, five, three and two lines, respectively, were chosen to grow on to produce T_2 grain for further analyses.

(1,3;1,4)-β-Glucan assay

Analysis of (1,3;1,4)-β-glucan content was performed using commercially available reagents (Megazyme International Ireland Ltd, Bray, Ireland), and a protocol based on [49] as detailed in Dimitroff et al. [40]. (1,3;1,4)-β-Glucan is reported as mg of (1,3;1,4)-β-glucan per mg of dry matter (% w/w).

Likewise, grain (1,3;1,4)-β-glucan content was assayed using commercially available reagents (Megazyme International Ireland Ltd, Bray, Ireland) following the protocol described in Ermawar et al. [34], using 15 mg ground flour. Aliquots of 100 μL were removed and stored at -20˚C for high performance anion-exchange chromatorgraphy (HPAEC) analysis of the oligosaccharides.

HPAEC analysis of DP3:DP4 content

The aliquots stored during the measurement of (1,3;1,4)-β-glucan were used to measure the ratio of the DP3 and DP4 oligosaccharides released during (1,3;1,4)-β-glucan hydrolysis. Samples were subjected to solid phase extraction [50] described in Ermawar et al. [34] on Varian Bond Elut Carbon 50 mg/1 mL columns, eluted with 55% acetonitrile, dried and re-suspended in 20 µl water. Each 20 µl sample had 10 µl 0.25 mM talose internal standard added and was derivatised as described in Comino et al. [51] with 0.5 M 1-phenyl-3-methyl-5-pyrazolone (PMP) in methanol. A 15 µl aliquot was run on an Aglient 1200 LC at 40˚C using a Phenomenex Kinetex XB-C18 2.6 µm 3x100 mm column as described in Comino et al. [51]. Peak integration and DP3:DP4 ratio calculation was performed by Agilent Chemstation software.

Transmission electron microscopy and immunocytochemistry

Detection of (1,3;1,4)-β-glucan in *N. benthamiana* tissue was carried out using fixation and embedding procedures described by Burton et al. [26] and with sectioning and imaging procedures described by Wilson et al. (2006). Here, a 1:500 dilution of the (1,3;1,4)-β-glucan-specific mouse primary antibody BG-1 (Biosupplies, Melbourne, Australia) [52] and a 1:30 dilution of Aurion goat IgG/IgM anti-mouse secondary antibody conjugated to 10 nm gold particles were used.

Phylogenetic analyses

For the CslE, CslG, CslJ and CslM families, the HMM for PF03552 was used with hmmalign [53] to assign residues to profile position for the *CslE, CslG, CslJ* and *CslM* families. The resulting alignment was stripped to assigned residues (> 0.6 posterior probability) only and alignments were manually inspected to evaluate their accuracy [2]. Models of sequence evolution were assessed for the final alignment using ModelOMatic [54] and jModelTest [55]. Best fit models were used to reconstruct phylogenetic trees of highest scoring models were reconstructed using the Bayesian MCMC package BEAST v2.3.1 [56] and the maximum-likelihood program RAxML8 [57].

BEAST analyses used a Yule tree model prior but uncorrelated log-normal, random and strict clock priors were run for at least 100,000,000 states or until stationarity was reached with all ESS values >200. Final BEAST trees were reconstructed using input alignments partitioned into the three separate codon positions, and substitution model parameters, including the rate heterogeneity model, base frequency and composition, and transition/transversion frequency, were allowed to vary across partitions. Convergence and mixing of the chain and the Effective Sample Sizes (ESS) of estimated parameters were monitored in TRACER v1.5 [58]. Stationarity was reached by 178 million generations. On this basis, the first 17.8 million sampled trees were discarded as burn-in. Treeannotator was used to find the maximum clade credibility tree and estimate posterior node support values.

Maximum-likelihood trees were reconstructed for three best fit models using the program RAxML8 [57], and the support for each model was assessed using gamma-based likelihood values. GTRGAMMAX using a codon partition (CP) scheme of CP1+CP2,CP3 was selected for the final maximum likelihood model. Three initial 1000 rapid bootstrap RAxML analyses were performed and the highest final gamma-based likelihood tree was used as the starting tree for 1000 rapid-hill climb searches and 1000 randomised ML tree searches. Notung [22] was used on the maximum clade credibility BEAST tree to infer gene duplication and loss events by reconciling the tree to a species tree created in PhyloT (http://phylot.biobyte.de/).

Non-synonymous to synonymous substitution ratio estimation

Non-synonymous to synonymous substitution ratios (*d*N:*d*S or ω) were calculated using the codeml program of the PAML 4.7 [24] package, and a specifically optimised version of codeml, slimcodeml [59]. The nw_utils package [60] was used to generate a subtree of the *CslJ* lineage and their eudicot sister clade. The branch-site model (M2a vs M1) was used to identify sites on the *CslJ* stem and crown group branches that have undergone positive selection (*d*N:*d*S > 1) against a null model with a fixed *d*N:*d*S ratio of 1.

Gene model analysis

GenePainter 2.0 [61] was used with full length *CslE, CslG, CslJ* and *CslM* sequences aligned using MUSCLE [62] to identify intron gain and loss throughout the evolution of these gene families.

Acknowledgements

The authors wish to acknowledge Ali Saleh Hassan for helping with the phylogenetic analysis and George Dimitroff for assistance with the heterologous expression. This work was supported by grants from the Australian Research Council and by postgraduate scholarships from the University of Adelaide and the Grains Research and Development Corporation to Julian Schwerdt.

References

1. Yin Y, Huang J, Xu Y. The cellulose synthase superfamily in fully sequenced plants and algae. BMC Plant Biol. 2009;9:99.

2. Schwerdt JG, MacKenzie K, Wright F, Oehme D, Wagner JM, Harvey AJ, Shirley NJ, Burton RA, Schreiber M, Halpin C, Zimmer J, Marshall DF, Waugh R, Fincher GB. Evolutionary Dynamics of the Cellulose Synthase Gene Superfamily in Grasses. Plant Physiol. 2015;168:968–83.

3. Cocuron J-C, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, Wilkerson CG. A gene from the cellulose synthase-like C family encodes a beta-1,4 glucan synthase. Proc. Natl. Acad. Sci. U. S. A. 2007;104:8550–5.

4. Liepman AH, Cavalier DM. The CELLULOSE SYNTHASE-LIKE A and CELLULOSE SYNTHASE-LIKE C families: recent advances and future perspectives. Front. Plant Sci. 2012;3:1–7.

5. Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM. Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. Proc. Natl. Acad. Sci. U. S. A. 1996;93:12637–42.

6. Popper ZA, Michel G, Hervé C, Domozych DS, Willats WGT, Tuohy MG, Kloareg B, Stengel DB. Evolution and diversity of plant cell walls: from algae to flowering plants. Annu. Rev. Plant Biol. 2011;62:567–90.

7. Popper ZA. Evolution and diversity of green plant cell walls. Curr. Opin. Plant Biol. 2008;11:286–92.

8. Doblin MS, Pettolino FA, Wilson SM, Campbell R, Burton RA, Fincher GB, Newbigin E, Bacic A. A barley cellulose synthase-like CSLH gene in transgenic Arabidopsis. 2009;106.

9. Lechat H, Amat M, Mazoyer J, Buléon A, Lahaye M. Structure and distrubution of glucomannan and sulfated glucan in the cell walls of the red alga Kappaphycus alvarezii (gigartinales, rhodophyta). J. Phycol. 2000;36:891–902.

10. Pettolino F, Sasaki I, Turbic A, Wilson SM, Bacic A, Hrmova M, Fincher GB. Hyphal cell walls from the plant pathogen *Rhynchosporium secalis* contain (1,3;1,6)-β-D-glucans, galacto- and rhamnomannans, (1,3;1,4)-β-D-glucans and chitin. FEBS J. 2009;276:3698–709.

11. Popper ZA. Primary Cell Wall Composition of Bryophytes and Charophytes. Ann. Bot. 2003;91:1–12.

12. Harris PJ. 11 Diversity in plant cell walls. Plant Divers. Evol. 2005;

13. Harris PJ, Hartley RD. Detection of bound ferulic acid in cell walls of the Gramineae by ultraviolet fluorescence microscopy. Nature. 1976;259:508–10.

14. Burton RA, Fincher GB. (1,3;1,4)-β-D-glucans in cell walls of the poaceae, lower plants, and fungi: A tale of two linkages. Mol. Plant. The Authors. All rights reserved.; 2009;2:873–82.

15. Fincher GB. Exploring the evolution of (1,3;1,4)-β-D-glucans in plant cell walls: comparative genomics can help! Curr. Opin. Plant Biol. 2009;12:140–7.

16. Burton RA, Wilson SM, Hrmova M, Harvey AJ, Shirley NJ, Medhurst A, Stone BA, Newbigin EJ, Bacic A, Fincher GB. Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1,3;1,4)-β-D-glucans. Science. 2006;311:1940–2.

17. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E, Zhang J, Huang L, Zhang L, Miao W, Zhang J, Ye Z, Miao C, Lin Z, Wang H, Zhou H, Yim WC, Priest HD, Zheng C, Woodhouse M, Edger PP. The pineapple genome and the evolution of CAM photosynthesis. Nat. Genet. 2015;47:1435–42.

18. Chaw SM, Chang CC, Chen HL, Li WH. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. 2004;58:424–41.

19. Farrokhi N, Burton RA, Brownfield L, Hrmova M, Wilson SM, Bacic A, Fincher GB. Plant cell wall biosynthesis: Genetic, biochemical and functional genomics approaches to the identification of key genes. Plant Biotechnol. J. 2006;4:145–67.

20. Lakatos L, Szittya G, Silhavy D, Burgyán J. Molecular mechanism of RNA silencing suppression mediated by p19 protein of tombusviruses. EMBO J. 2004;23:876–84.

21. Hazen SP, Scott-Craig JS, Walton JD. Cellulose synthase-like (CSL) genes of rice Cellulose Synthase-Like Genes of Rice 1. 2002;

22. Liebert MA, Chen K, Durand D, Farach-Colton M, Al CET. NOTUNG : A Program for Dating Gene Duplications. 2000;7:429–47.

23. Hammesfahr B, Odronitz F, Mühlhausen S, Waack S, Kollmar M. GenePainter: a fast tool for aligning gene structures of eukaryotic protein families, visualizing the alignments and mapping gene structures onto protein structures. BMC Bioinformatics. 2013;14:77–77.

24. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 2007;24:1586–91.

25. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2010;gkq1019.

26. Burton RA, Collins HM, Kibble NA, Smith JA, Shirley NJ, Jobling SA, Henderson M, Singh RR, Pettolino F, Wilson SM, Bird AR, Topping DL, Bacic A, Fincher GB. Over-expression of specific HvCslF cellulose synthase-like genes in transgenic barley increases the levels of cell wall (1,3;1,4)-β-D-glucans and alters their fine structure. Plant Biotechnol. J. 2011;9: 117–35.

27. Fry SC, Nesselrode BHW a, Miller JG, Mewburn BR. Mixed-linkage (1→3,1→4)-beta-D-glucan is a major hemicellulose of Equisetum (horsetail) cell walls. New Phytol. 2008; 179:104–15.

28. Nobles DR, Brown RM. The pivotal role of cyanobacteria in the evolution of cellulose synthases and cellulose synthase-like proteins. Cellulose. 2004; 11:437–48.

29. Nemeth C, Freeman J, Jones HD, Sparks C, Pellny TK, Wilkinson MD, Dunwell J, Annica AM, Andersson PA, Guillon F, Saulnier L, Mitchell RAC, Shewry PR. Down-regulation of the CSLF6 gene results in decreased (1,3;1,4)-β-D-glucan in endosperm of wheat. Plant Physiol. 2010; 152:1209–18.

30. Taketa S, Yuo T, Tonooka T, Tsumuraya Y, Inagaki Y, Haruyama N. Functional characterization of barley betaglucanless mutants demonstrates a unique role for CslF6 in (1,3;1,4) -b- D -glucan biosynthesis. 2012; 63:381–92.

31. Vega-Sanchez ME, Verhertbruggen Y, Christensen U, Chen X, Sharma V, Varanasi P, Jobling SA. Loss of Cellulose synthase-like F6 function affects mixed-linkage glucan deposition, cell wall mechanical properties, and defense responses in vegetative tissues of rice. Plant Physiol. 2012;159:56–69.

32. Hu G, Burton C, Hong Z, Jackson E. A mutation of the cellulose-synthase-like (CslF6) gene in barley (Hordeum vulgare L.) partially affects the β-glucan content in grains. J. Cereal Sci; 2014;59:189–95.

33. Burton RA, Jobling S a, Harvey AJ, Shirley NJ, Mather DE, Bacic A, Fincher GB. The genetics and transcriptional profiles of the cellulose synthase-like HvCslF gene family in barley. Plant Physiol. 2008;146:1821–33.

34. Ermawar RA, Collins HM, Byrt CS, Betts NS, Henderson M, Shirley NJ, Schwerdt JG, Lahnstein J, Fincher GB, Burton RA. Distribution, structure and biosynthetic gene families of (1,3;1,4)-β-glucan in Sorghum bicolor. J. Integr. Plant Biol. 2015;57:429–45.

35. Christensen U, Alonso-Simon A, Scheller H V., Willats WGT, Harholt J. Characterization of the primary cell walls of seedlings of Brachypodium distachyon — A potential model plant for temperate grasses. Phytochemistry. 2010;71:62–9.

36. Kondrashov F a, Rogozin IB, Wolf YI, Koonin E V. Selection in the evolution of gene duplications. Genome Biol. 2002;3:RESEARCH0008.

37. Francino MP. An adaptive radiation model for the origin of new gene functions. Nat. Genet. 2005;37:573–8.

38. Hughes AL. The evolution of functionally novel proteins after gene duplication. Proc. Biol. Sci. 1994;256:119–24.

39. Richmond TA, Somerville CR. The cellulose synthase superfamily. Plant Physiol. 2000;124:495–8.

40. Dimitroff G, Little A, Lahnstein J, Schwerdt JG, Srivastava V, Bulone V, Burton RA, Fincher GB. (1,3;1,4)-β-Glucan biosynthesis by the CSLF6 enzyme: Position and flexibility of catalytic residues influence product fine structure. Biochemistry (Mosc.). 2016;55:2054-2061

41. Burton R a, Shirley NJ, King BJ, Harvey AJ, Fincher GB. The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. Plant Physiol. 2004;134:224–36.

42. Burton RA, Jobling S a, Harvey AJ, Shirley NJ, Mather DE, Bacic A, Fincher GB. The genetics and transcriptional profiles of the cellulose synthase-like HvCslF gene family in barley. Plant Physiol. 2008;146:1821–33.

43. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by

geometric averaging of multiple internal control genes. Genome Biol. 2002;3:RESEARCH0034.

44. Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu S-H, Jiang N, Robin BC. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J. 2012;71:492-502.

45. Vickers CE, Xue G, Gresshoff PM. A novel cis-acting element, ESP, contributes to high-level endosperm-specific expression in an oat globulin promoter. Plant Mol. Biol. 2006;62:195–214.

46. Burton R a, Gidley MJ, Fincher GB. Heterogeneity in the chemistry, structure and function of plant cell walls. Nat. Chem. Biol.; 2010;6:724–32.

47. Sainsbury F, Thuenemann EC, Lomonossoff GP. PEAQ: Versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. Plant Biotechnol. J. 2009;7:682–93.

48. Hofgen R, Willmitzer L. Storage of competent cells for Agrobacterium transformation. Nucleic Acids Res. 1988;16:9877–9877.

49. McCleary B V, Codd R. Measurement of (1-3),(1-4)- β-D-Glucan in barley and oats: a streamlined enzymic procedure. J. Sci. Food Agric. 1991;55:303–12.

50. Redmond JW, Packer NH. The use of solid-phase extraction with graphitised carbon for the fractionation and purification of sugars. Carbohydr. Res. 1999;319:74–9.

51. Comino P, Shelat K, Collins H, Lahnstein J, Gidley MJ. Separation and purification of soluble polymers and cell wall fractions from wheat, rye and hull less

barley endosperm flours for structure-nutrition studies. J. Agric. Food Chem. 2013;61:12111–22.

52. Meikle PJ, Hoogenraad NJ, Bonig I, Clarke AE, Stone B a. A (1→3,1→4)-β-glucan-specific monoclonal antibody and its use in the quantitation and immunocytochemical location of (1→3,1→4)-β-glucans. Plant J. Cell Mol. Biol. 1994;5:1–9.

53. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39:W29-37.

54. Whelan S, Allen JE, Blackburne BP, Talavera D. ModelOMatic: Fast and Automated Model Selection between RY, Nucleotide, Amino Acid, and Codon Substitution Models. Syst. Biol. 2014;64:42–55.

55. Posada D. jModelTest: Phylogenetic Model Averaging. Mol. Biol. Evol. 2008;25:1253–6.

56. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 2012;29:1969–73.

57. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

58. Drummond A, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 2007;7:214.

59. Schabauer H, Valle M, Pacher C, Stockinger H, Stamatakis A, Robinson-Rechavi M, Yang Z, Salamin N. SlimCodeML: An Optimized Version of CodeML for the Branch-Site Model. 2012 IEEE 26th Int. Parallel Distrib. Process. Symp. Workshop PhD Forum. Ieee; 2012;706–14.

60. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26:1669–70.

61. Mühlhausen S, Hellkamp M, Kollmar M. GenePainter v. 2.0 resolves the taxonomic distribution of intron positions. Bioinformatics. 2014;btu798.

62. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

63.The International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491:711–6.

Figures

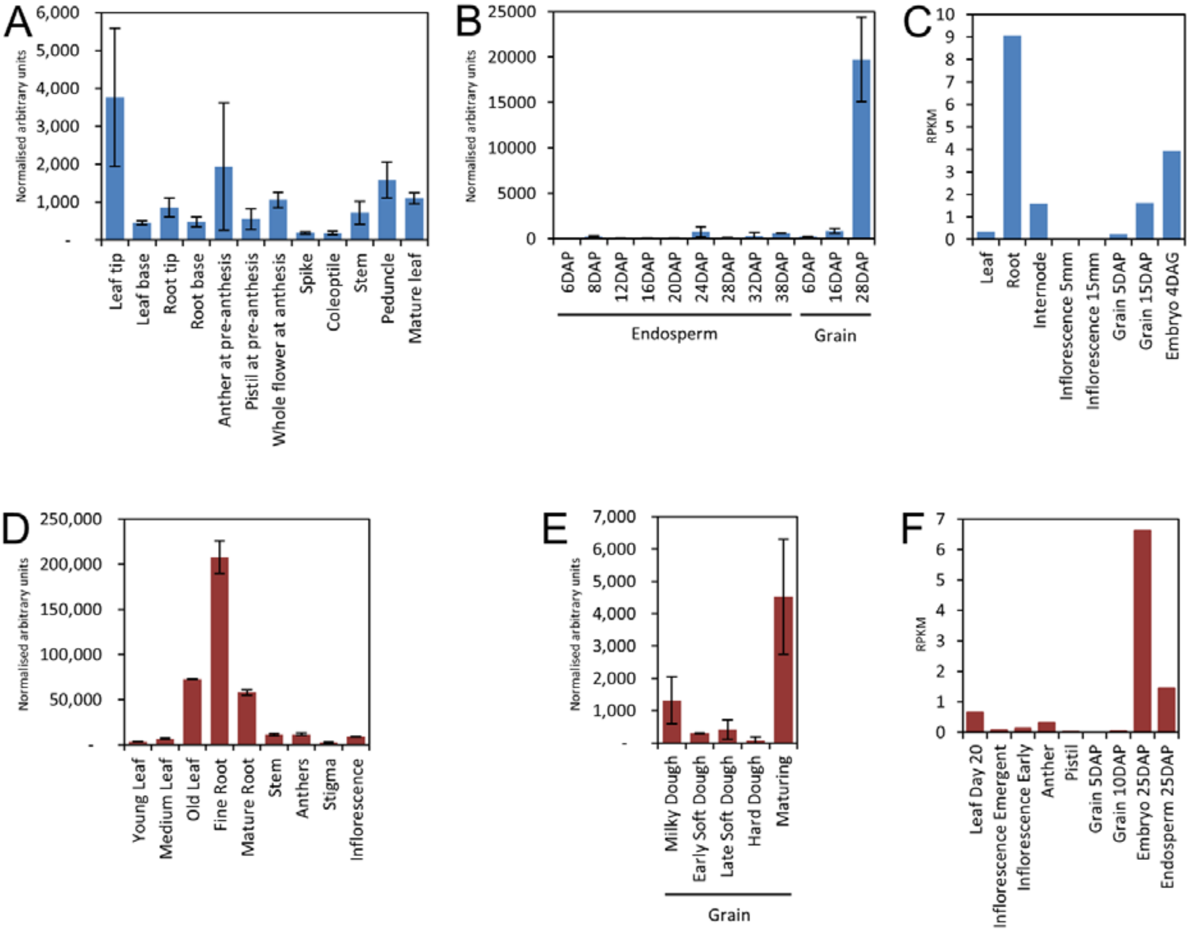


Figure 1. QPCR and RNAseq transcript profiles for *HvCslJ1* (Blue) and *SbCslJ1* (Red) genes. (A) *HvCslJ1* QPCR transcript levels from cDNA sets described in Burton et al. (2008) [42]. (B) *HvCslJ1* QPCR transcript levels from endosperm and whole grain samples (see methods and materials). (C) *HvCslJ1* transcript levels from an RNAseq dataset described in Consortium, T. I. B. G. S. [63]. (D/E) *SbCslJ1* QPCR transcript levels from cDNA sets described in Ermawar et al. [34]. (F) *SbCslJ1* transcript levels from an RNAseq dataset described in Davidson et al. [44]. Error bars represent the standard error. DAP, days after pollination, DAG, days after germination

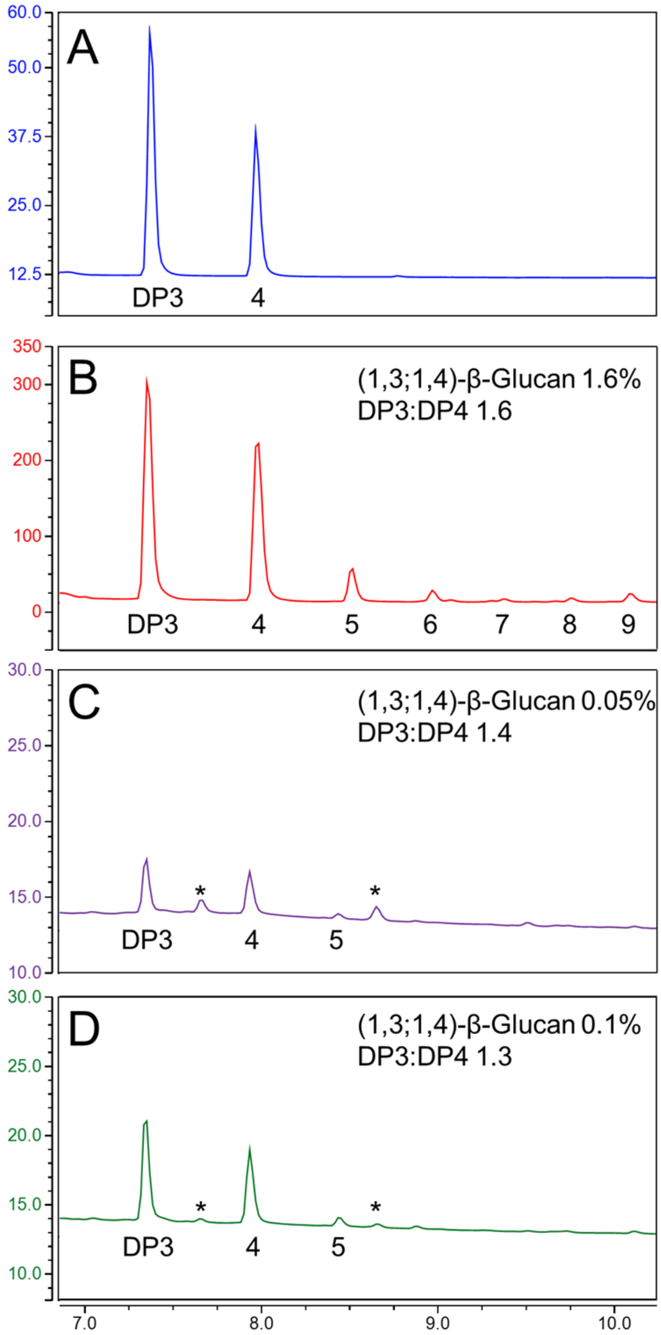


Figure 2. (1,3;1,4)-β-Glucan amounts (%,w/w) and fine structure (DP3:DP4) produced by heterologous transient expression. HPAEC traces of DP3 and DP4 oligosaccharide standards generated from commercial barley (1,3;1,4)-β-glucan (A) and oligosaccharides released post lichenase digestion from *N. benthamiana* leaf tissue expressing *HvCslF6* (B), *HvCslH1* (C) and *HvCslJ1* (D).*, unknown compounds, X axis: minutes, Y axis: nanocoulombs (nC).

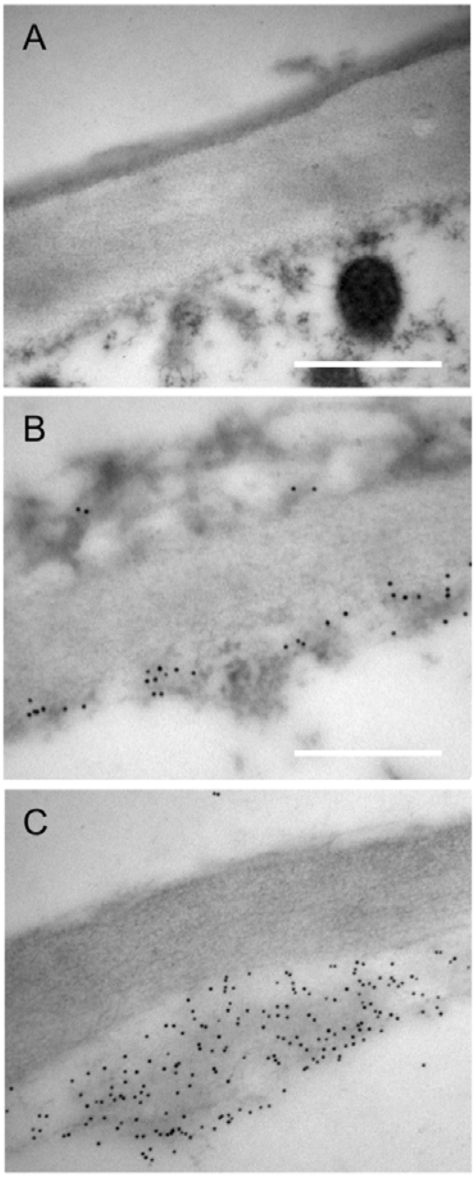


Figure 3. Transmission electron micrograph labelled with the (1,3;1,4)-β-glucan specific antibody. Immunogold labeling of *N. benthamiana* leaf tissue expressing an empty vector control (A), *HvCslJ1* (B) and *HvCslF6* (C). Scale bars = 1 µm.

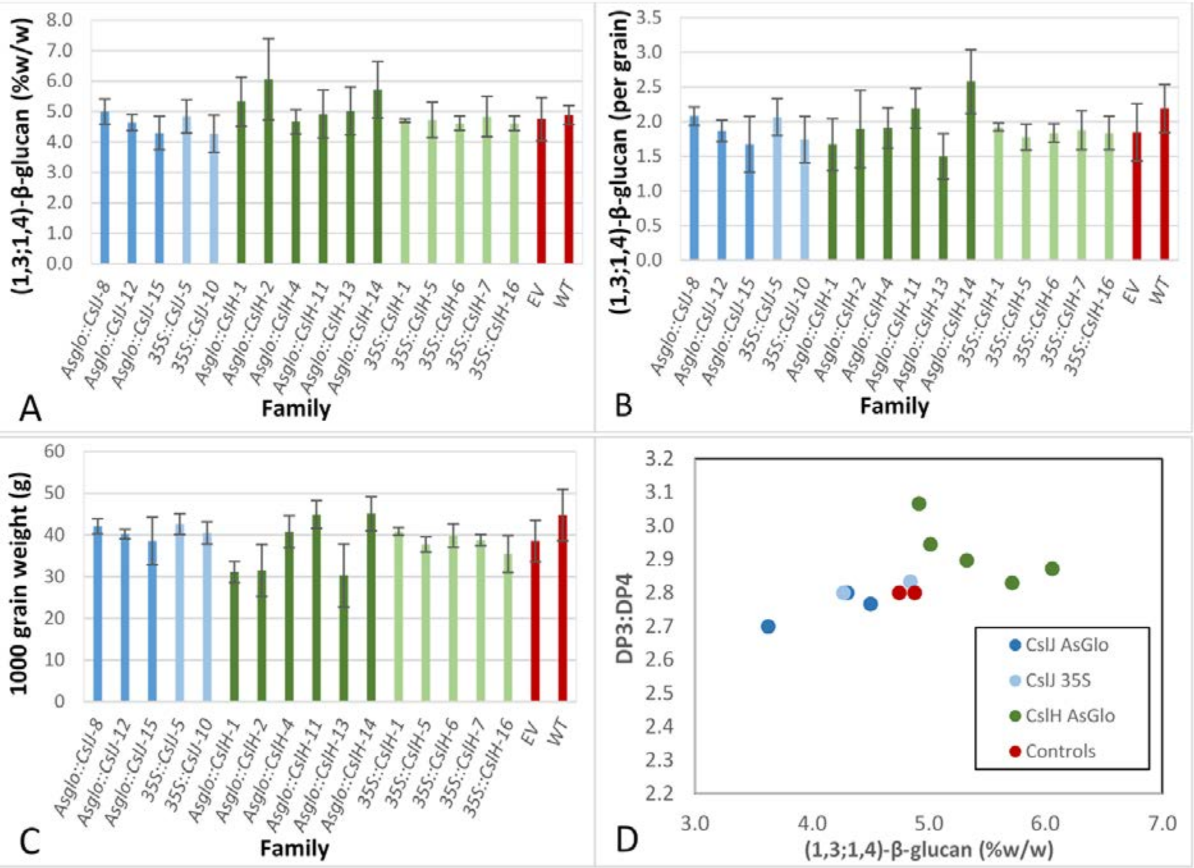


Figure 4. Analysis of transgenic grain. (A) (1,3;1,4)-β-glucan amount (%w/w) (B) (1,3;1,4)-β-glucan amount per grain (C) grain weight (g per 1000 grain) and (D) fine structure (DP3:DP4 vs (1,3;1,4)-β-Glucan) produced by the overexpression of barley cellulose synthase-like genes**.** Blue = *HvCslJ1*, Green = *HvCslH1*, Red = controls, WT = wild type barley, EV = empty vector. Error bars represent standard deviations.

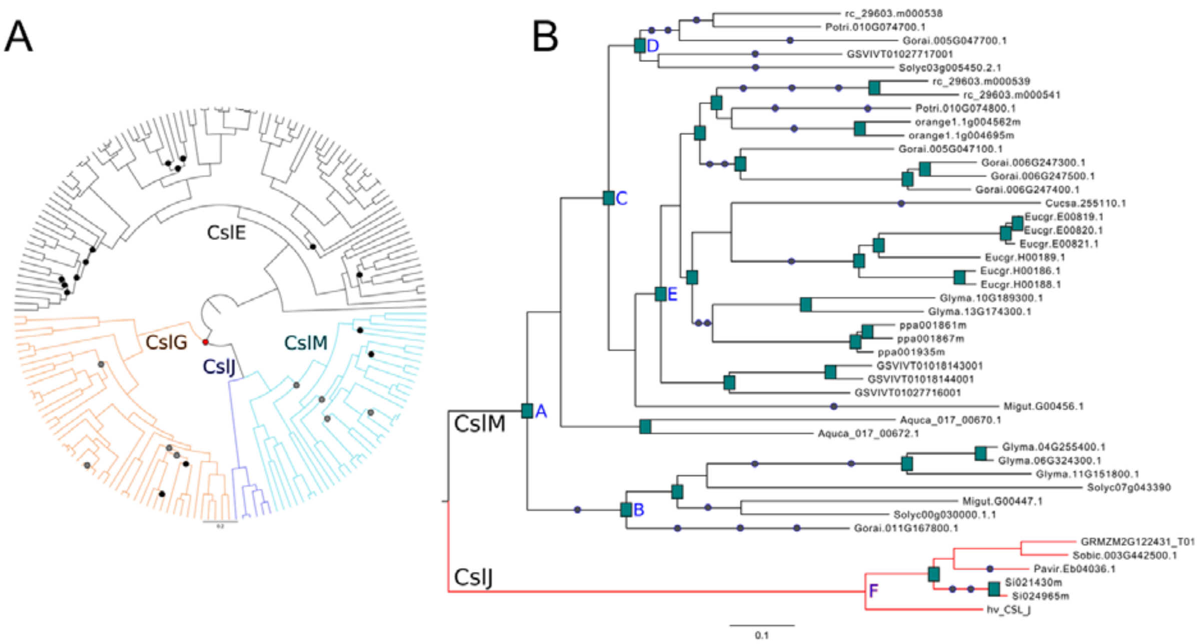


Figure 5. Bayesian maximum clade credibility trees. (A) BEAST maximum clade credibility tree of 21 fully sequenced eudicot and monocot *CslE, CslG, CslJ* and *CslM* genes. Node support values (posterior probabilities) < 0.95 and > 0.85 are shown as grey dots, < 0.85 and > 0.50 are shown as black dots, < 0.5 are shown as red dots. (B) Subtree of *CslJ* and *CslM* with gene duplications annotated at cyan squares on nodes and gene losses annotated as blue circles on branches. Major lineages under positive selection are coloured red.

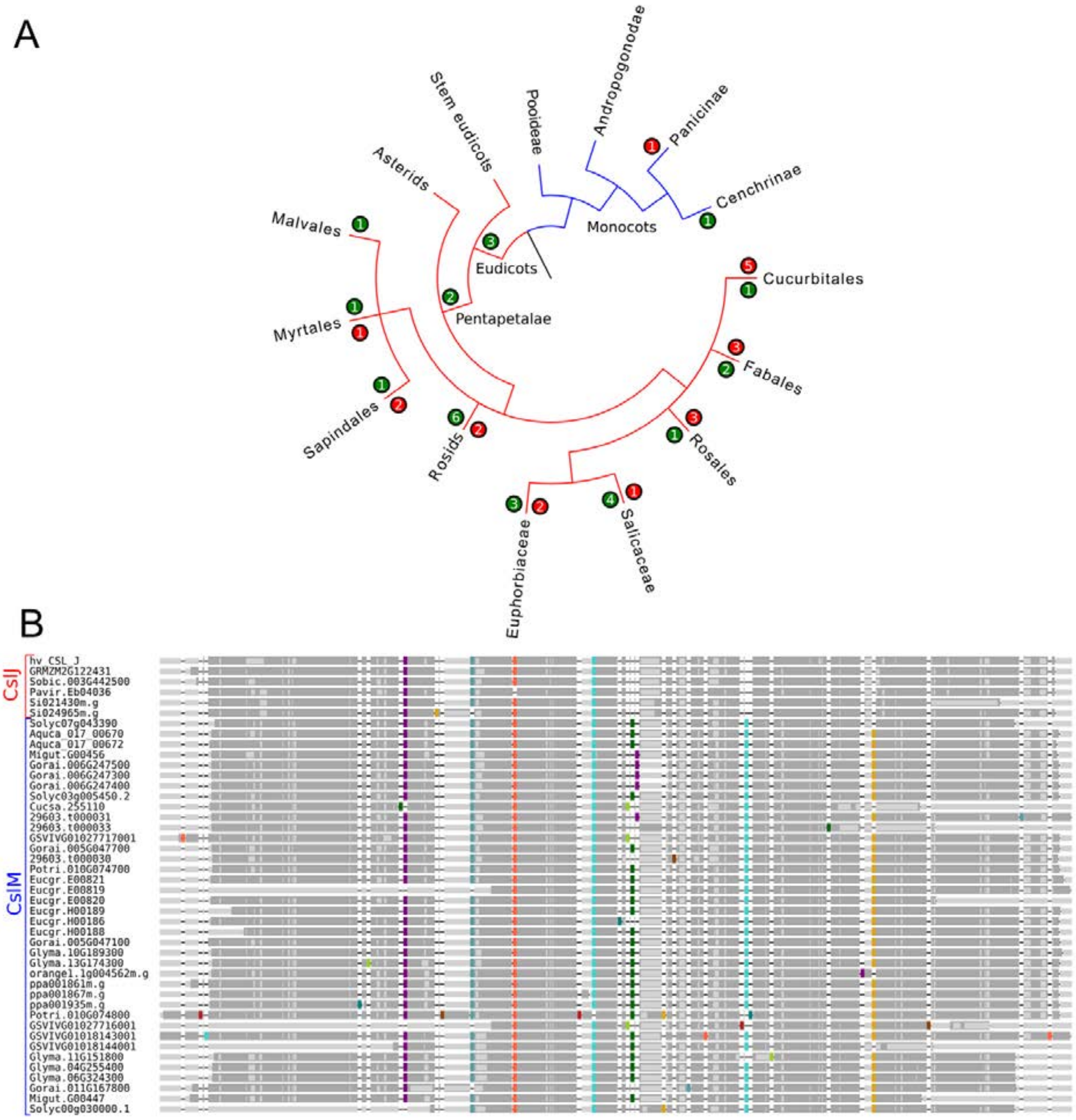


Figure 6. *CslJ* and *CslM* gene structure. (A) Sampled plant species tree indicates number of gained introns with green circles and number of lost introns with red circles. (B) *CslJ* and *CslM* alignment identifies conserved intron positions with coloured markers.

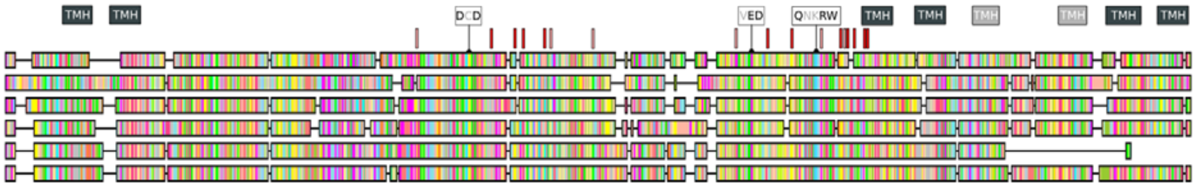
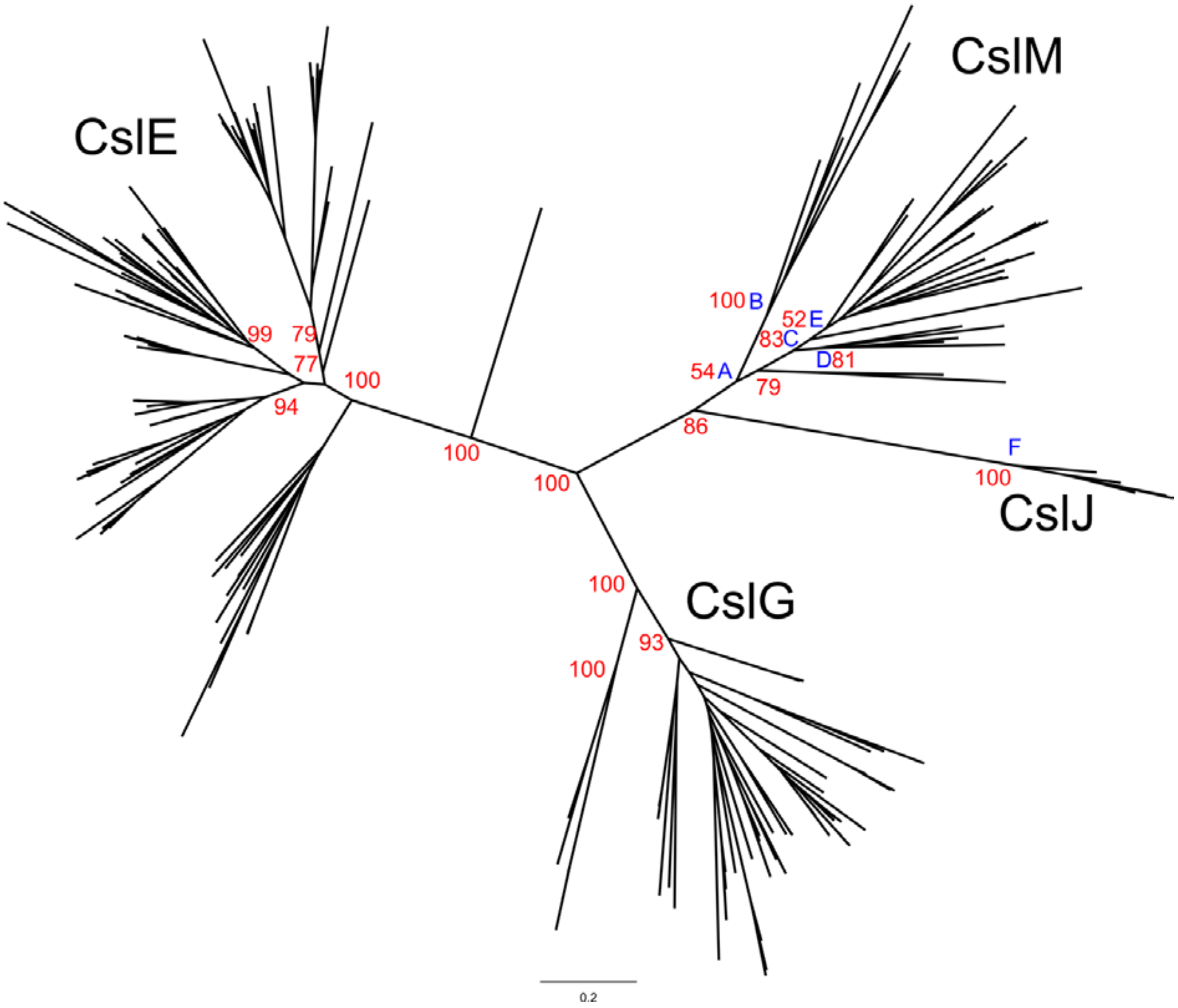
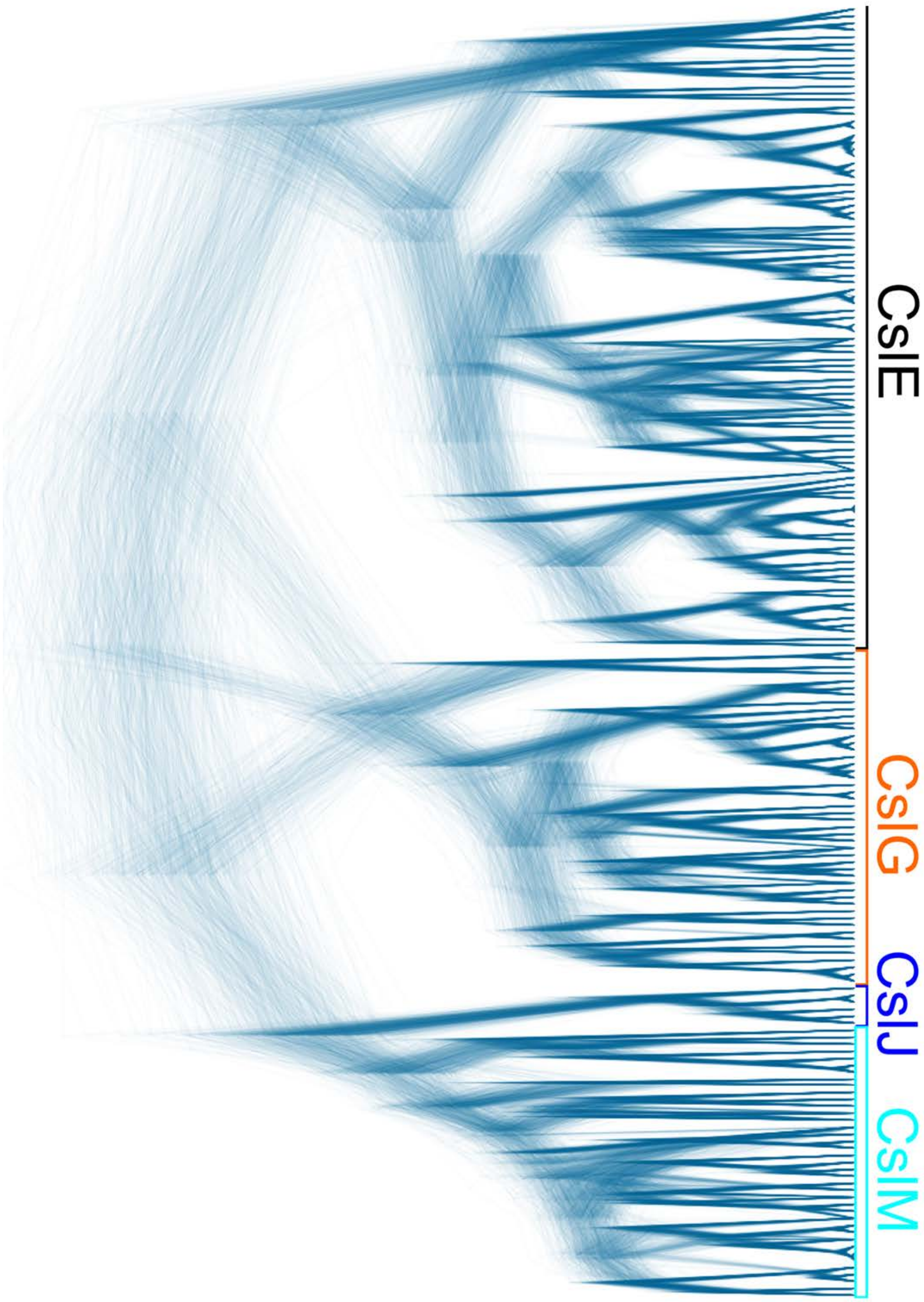


Figure 7. Alignment of *CslJ* sequences. Amino acid colour coded as default in Geneious 8.1.5 (http://www.geneious.com). Predicted transmembrane helices are annotated with grey boxes. Core catalytic motifs (DxxD; TED; QxxRW) are annotated using white boxes. Positions of amino acids under selection with a posterior probability of > 0.95 are indicated with red bars. Amino acids with a > 0.85 posterior probability are indicated with pink bars above the alignment.



Supplementary Figure 1. Best known maximum likelihood tree for *CslE, CslG, CslJ* and *CslM* sequences from 21 fully sequenced eudicot and monocot species. Bootstrap support values are annotated on deep nodes in red. Labeled clades from Figure 5A are annotated with blue letters.



Supplementary Figure 2. DensiTree plot visualising trees sampled (at a frequency of 100,000 trees) in the Bayesian analysis.

Family	No	Grain weight (g per 1000 grain)			(1,3;1,4)-β-Glucan (%w/w)			(1,3;1,4)-β-Glucan (per grain)			DP3:DP4			
		*Mean	Range	sd	*Mean	Range	sd	*Mean	Range	sd	No	$^{*\#}$Mean	Range	sd
*As*GLO::CslJ-8	5	42.1^{ab}	39.6-43.9	1.8	5.0^{ac}	4.5-5.4	0.4	2.1^{bcd}	2.0-2.2	0.1	3	2.8^{a}	2.7-2.8	0.06
*As*GLO::CslJ-12	5	40.2^{bcd}	39.2-41.5	1.2	4.6^{ab}	4.3-5.0	0.3	1.9^{abc}	1.7-2.1	0.2	1	2.8	2.9	
*As*GLO::CslJ-15	4	38.6^{bd}	30.2-42.9	5.7	4.3^{a}	3.6-4.9	0.5	1.7^{ab}	1.1-2.0	0.4	1	2.7	2.7	
35S::CslJ-5	5	42.6^{ab}	39.6-45.8	2.5	4.8^{abc}	3.9-5.4	0.5	2.1^{bc}	1.6-2.3	0.3	3	2.8^{a}	2.8-2.9	0.06
35S::CslJ-10	3	40.5^{abd}	36.0-42.6	2.6	4.3^{a}	3.8-5.3	0.6	1.7^{abe}	1.4-2.3	0.3	1	2.8	2.8	
*As*GLO::CslH-1	4	31.1^{ed}	28.0-34.2	2.5	5.3^{bcd}	4.2-6.0	0.8	1.7^{ab}	1.2-2.1	0.4	2	2.9^{a}	2.8-3.0	0.11
*As*GLO::CslH-2	6	31.5^{ed}	19.6-38.3	6.3	6.1^{d}	3.9-7.5	1.3	1.9^{abc}	1.3-2.9	0.6	3	2.9^{a}	2.8-3.0	0.11
*As*GLO::CslH-4	3	40.8^{abd}	37.6-45.1	3.9	4.7^{abc}	4.2-5.0	0.4	1.9^{abc}	1.6-2.2	0.3				
*As*GLO::CslH-11	5	44.9^{ac}	41.3-49.6	3.3	4.9^{ac}	3.7-5.6	0.8	2.2^{cde}	1.8-2.5	0.3	2	3.1^{b}	3.1-3.1	0.02
*As*GLO::CslH-13	3	30.3^{ed}	22.5-37.5	7.5	5.0^{ac}	4.2-5.8	0.8	1.5^{a}	1.3-1.9	0.3	1	2.9	2.9	
*As*GLO::CslH-14	4	45.1^{a}	43.0-51.2	4.1	5.7^{cd}	4.8-7.0	0.9	2.6^{d}	2.1-3.0	0.5	3	2.8^{a}	2.8-2.8	0.03
35S::CslH-1	5	40.9^{abd}	39.9-41.6	0.9	4.7^{abc}	4.6-4.7	0.0	1.9^{abc}	1.9-2.0	0.1				
35S::CslH-5	4	37.7^{bd}	35.7-40.1	1.8	4.7^{ab}	4.0-5.4	0.6	1.8^{abc}	1.5-1.9	0.2				
35S::CslH-6	4	39.8^{abd}	37.2-42.4	2.8	4.6^{ab}	4.4-4.9	0.2	1.8^{abc}	1.7-2.0	0.1				
35S::CslH-7	4	38.8^{bd}	37.2-40.3	1.4	4.8^{abc}	4.1-5.5	0.7	1.9^{abc}	1.6-2.2	0.3				
35S::CslH-16	4	35.5^{d}	29.7-40.5	4.4	4.6^{ab}	4.4-5.0	0.2	1.8^{ab}	1.4-1.9	0.2				
Empty Vector	13	38.6^{bd}	26.3-45.1	5.0	4.7^{ab}	3.8-6.3	0.7	1.8^{ab}	1.0-2.6	0.4	2	2.8^{a}	2.8-2.8	0.00
Wild type	9	44.8^{ac}	32.2-51.2	6.1	4.9^{abc}	4.4-5.5	0.3	2.2^{cd}	1.6-2.6	0.3	1	2.8	2.8	

*Mean results with letters in common are not significantly different (P<0.05), $^{\#}$ no letter: not enough replicates to analyse. Red: significantly different to the wild type, blue: significantly different to the empty vector and purple significantly different to both the wild type and empty vector

Supplementary Table S1 Mean, standard deviation and range in grain weight, (1,3;1,4)-β-Glucan amounts and DP3:DP4 produced by the overexpression of barley cellulose synthase-like genes in stable transgenic barley endosperm (T_2).

Supplementary Table S2: Amino acid sites under selection in the *CslJ* lineage

Alignment position	HvCslJ position	Amino acid	Posterior probability (BEB)
1	241	M	0.878
53	293	S	0.972
67	308	R	0.996
69	310	D	0.985
85	326	S	0.957
89	330	A	0.89
104	359	A	0.868
124	435	E	0.886
128	456	C	0.973
142	473	P	0.954
165	496	M	0.886
174	505	C	0.956
175	508	P	0.827
177	510	A	0.916
178	511	S	0.992
179	512	A	0.988
180	513	A	0.973
188	519	G	0.993
189	520	F	0.999

Statement of Authorship

Title of Paper	Phylogenetic analysis of the angiosperm cellulose synthase superfamily: recommendations for a standardised nomenclature
Publication Status	□ Published □ Accepted for Publication □ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Julian Schwerdt		
Contribution to the Paper	Performed all analyses and prepared manuscript.		
Overall percentage (%)	90		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	2/8/16

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Neil Shirley		
Contribution to the Paper	Assisted with data collection and curation.		
Signature		Date	2/8/2016

Name of Co-Author	Rachel Burton		
Contribution to the Paper	Conceived project and designed experiments.		
Signature		Date	2/8/16

Name of Co-Author	Geoffrey Fincher		
Contribution to the Paper	Conceived project and designed experiments.		
Signature		Date	2/8/16.

Phylogenetic analysis of the angiosperm cellulose synthase superfamily: recommendations for a standardised nomenclature

Julian Schwerdt, Neil Shirley, Rachel Burton and Geoff Fincher

Abstract

The plant cellulose synthase (*CesA*) and cellulose synthase-like (*Csl*) genes comprise a large and functionally important superfamily that is a major focus of cell wall research. However, the lack of any standardised and consistent system for classifying and identifying *CesA* and *Csl* genes in different species currently hinders communication of the many varied studies of their biochemical activities and functional roles in a wide range of plant species. In the present paper, we address this need using model-based analyses to reconstruct phylogenetic relationships and infer duplication events among the *CesA* and *Csl* genes of 22 fully sequenced flowering plant genomes. The recovered phylogenetic history and identified discriminatory protein motifs are used to construct a system for naming new existing *CesA* and *Csl* genes. We retain the use of *CesA* and *Csl* root symbols for cellulose synthase and cellulose synthase-like gene families, respectively, and existing subfamily designation. Gene families are numbered where lineages existed prior to the monocot-eudicot divergence; major subsequent duplications are differentiated by a character suffix.

Introduction

The plant cell wall contains a diverse range of polysaccharides including cellulose, xyloglucan, (1,3;1,4)-β-glucan, heteroxylan, heteromannan and pectin. Polysaccharide presence and abundance varies substantially across plant lineages, with eudicots, for instance, displaying a greater abundance of xyloglucan and pectic polysaccharides compared with monocots, which contain proportionally more heteroxylans (Scheible and Pauly, 2004). The cellulose synthase genes CesA1 and CesA2 were initially characterised from cotton fibres. Subsequent sequence database homology searches systematically identified six gene families that were designated cellulose synthase-like (Csl) genes: CslA, CslB, CslC, CslD, CslE and CslG (Richmond and Somerville, 2000). As more plant species were sequenced a further four Csl families were identified, namely CslF, CslH (Hazen et al., 2002), CslJ (Fincher, 2009) and CslM (Schwerdt et al., in preparation).

Functional characterisation of the cellulose synthase superfamily has proven challenging. Some *Csl* families have been associated with the synthesis of particular polysaccharides, notably *CslF, CslH* and *CslJ* with (1,3;1,4)-β-glucan (Burton et al., 2006; Doblin et al., 2009; Schwerdt et al., in prep), *CslC* with xyloglucan (Cocuron et al., 2007; Dwivany et al., 2009), and *CslA* with mannan (Liepman et al., 2005). However, it is unclear whether a one-to-one relationship exists between *Csl* subfamilies and particular polysaccharides, and at least some clades are likely to be involved in the production of multiple polymers (Dwivany et al., 2009). Resolving the functional roles of this large and complex gene superfamily is an active area of research (Scheible and Pauly, 2004; Fincher, 2009). However, the absence of any standardised and unambiguous nomenclature for classifying *Csl* genes makes it difficult to compare the results from different research groups. Relationships among the cellulose synthase and cellulose synthase-like families remain unresolved (Yin et al., 2009), and most genes are numbered according to the order in which they were discovered, or are named by homology to the two model systems, *Arabidopsis thaliana* and *Oryza sativa*.

To address this problem we recommend here a revised and unifying nomenclature based on detailed phylogenetic analysis of 22 fully sequenced angiosperms. The new classification system uses well-resolved phylogenetic relationships, inferred orders of duplication events, and identified discriminatory protein motifs. We expect that adoption of this system will reduce confusion in the scientific literature, for instance by removing bias towards the *Arabidopsis thaliana* and *Oryza sativa* model systems where differences in gene duplication and loss complicate a numerical nomenclature. Additionally, we expect the system to prove extensible to accommodate future discoveries relating to the functional classification of specific clades in the superfamily.

Materials and Methods

Data sources

A total of 741 candidate loci and 8kb of upstream and downstream flanking sequences were retrieved from Phytozome 10 (Goodstein et al., 2012), and the presence of PFAM PF03552 Cellulose_synt was used to identify *CesA, CslB, CslD, CslE, CslF, CslG, CslH, CslJ* and *CslM* candidate sequences. The *CslA* and *CslC* families were retrieved using the Glycosyl transferase family GT2 PF00535 domain. In addition, we ran BLAST and HMMsearch (Camacho et al., 2009; Finn et al., 2011) locally against the downloaded complete protein datasets from Phytozome to verify candidates and identify any missing putative cellulose synthase superfamily sequences. Barley sequence data were sourced from the International Barley Sequencing Consortium assembly (Mayer et al., 2012). The barley putative sequences were identified using BLAST and HMMsearch (Camacho et al., 2009; Finn et al., 2011) with the PF03552 and PF00535 profiles.

Characterizations and Quality control

In order to mitigate annotation and sequence error, we implemented an annotation pipeline with a final manual curation stage. We first reconstructed a neighbour joining tree of a selected cellulose synthase superfamily representative dataset and the predicted transcript using PHYLIP (Felsenstein, 1993) and a HKY substitution model, to identify likely clade membership. We then used the FGENESH+ (Solovyev, 2002) gene predictor to generate alternate transcripts using manually curated characteristic clade member protein sequences as homologues. Each predicted transcript was annotated for splice site junctions using Spidey (Wheelan et al., 2001), and functional domains with InterProScan (Jones et al., 2014). Every FGENESH+ transcript for each locus was translated and aligned with appropriate subfamily members using MUSCLE (Edgar, 2004), and back-translated to the original nucleotide sequence. Each alignment was manually inspected to maximize accuracy of each locus' gene model.

Multiple sequence alignment

The hidden Markov model (HMM) for PF03552 was used with hmmalign (Finn et al., 2011) to assign amino acid residues to profile position for *CslB/D/E/F/G/H/J/M/CesA*. HMM sites with assignments below an average 0.6 posterior probability were manually stripped from the alignments to produce final versions. Because the *CslC* gene family does not contain the large 5` conserved domains present in *CslA*, the PF00535 HMM is short relative to the *CslA* average sequence length. For this reason, Clustal Omega (Sievers et al., 2011) was used with mBed distance clustering and the PF00535 HMM as background data to align the *CslA* and *CslC* families. BMGE (Criscuolo and Gribaldo, 2010) was used to remove mis-aligned and ambiguous sites. The *CslA/CslC* alignment nucleotide dataset was constructed by mapping codons back onto the amino acid alignments. Subfamily alignments used in this work are subsets of these major datasets.

Substitution model selection

Best fit models of sequence evolution for final alignments were selected using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) implemented in jModelTest, PartitionFinder, ModelOMatic and ProtTest (Posada, 2008; Darriba et al., 2011; Lanfear et al., 2012; Whelan et al., 2014). Additionally, final gamma-based likelihood values from RAxML (Stamatakis, 2014) were compared for the GTRGAMMA, GTRGAMMAX, GTRCAT, GTRCATX, PROTGTRGAMMA, PROTGTRCAT, PROGTRGAMMAX, PROGTRCATX, PROTLGGAMMA, PROTLGGAMMAX, PROTLGCAT, PROTLGCATX, PROJTTGAMMA, PROTJTTGAMMAX, PROTJTTCAT and PROJTTCATX substitution models. Nucleotide models were tested using two codon position (CP) partitioning schemes:

CP1,CP2,CP3 and CP1+CP2,CP3. Models of sequence evolution for Bayesian analyses conducted in BEAST2 (Bouckaert et al., 2014) were assessed by comparing convergence and effective sample sizes (ESS) for important parameters including likelihood, posterior, ucld.mean and coefficient of variation using Tracer 1.5 (Drummond and Rambaut, 2007). The GTR+I+G, GTR+G, HKY+G, HKY+I+G nucleotide substitution models were tested for the same partition schemes used for the maximum likelihood tests. Amino acid substitution models tested included LG, WAG and JTT. Additionally, strict, uncorrelated log normal and random clock models were assessed for each substitution model.

Phylogenetic analyses

Phylogenies of PF00535 and PF03552 sequences and nested subfamilies were reconstructed using amino acid and nucleotide data, under both maximum likelihood (ML) using RAxML version 8.1.2 (Stamatakis, 2014) and Bayesian inference using the MCMC package BEAST 2.2.1 (Bouckaert et al., 2014). Nucleotide RAxML analyses used a CP1+CP2,CP3 codon position partitioning scheme and the GTRGAMMAX substitution model because these parameters produced the highest likelihood scores during model assessment for all data. The PROTGTRGAMMAX amino acid substitution model produced the highest likelihood for all data and out-performed best fit empirical models. All RAxML analyses began with a 1000 rapid bootstrap analysis and 200 best ML tree search. The tree with the highest likelihood was used as the starting tree for 1000 rapid hill-climb ML tree searches and 1000 randomised tree searches. This procedure was repeated three times for each dataset and the trees with the highest final gamma-based likelihood were chosen. ML phylogenies were unrooted because of the difficulty in assigning an appropriate outgroup for an entire gene family that is at least as old as the first embryophyte.

Bayesian analyses were conducted using BEAST 2.2.1 (Bouckaert et al., 2014). Nucleotide alignments were partitioned into the three codon positions, and each partition was unlinked, that is, substitution model parameters, rate heterogeneity model and base frequency were free to vary across codon positions. Inspection of the coefficient of variation parameter indicated that the relaxed-clock (log-normal distribution of nucleotide rate variation) was appropriate for our data. The GTR substitution model with no gamma rate variation or invariant sites and a Yule tree prior was used for each nucleotide dataset. Amino acid alignments were performed using the LG with gamma rate variation substitution model, an uncorrelated log-normal relaxed clock and Yule tree prior. Convergence was monitored in TRACER v1.5 (Drummond and Rambaut, 2007) by assessing the effective sample sizes (ESS), posterior probabilities and likelihood values of the model parameters. Each analysis was repeated three times and was run for at least 50,000,000 states, logging every 1000, or until stationarity was reached. Unlike RAxML, BEAST determines the root position from the data so that rooted trees are generated without assigning an outgroup. A species tree was obtained from previous work (Lee et al., 2011) and PhyloT (http://phylot.biobyte.de/). Notung (Liebert et al., 2000) was used to reconcile the BEAST and RAxML gene trees with the species trees to identify gene duplication and loss.

Motif discovery

We developed BioPerl (bioperl.org) code to extract each selected node’s full length representative sequences and a background set composed of all others. We performed a discriminatory motif discovery with DEME (Redhead and Bailey, 2007) on each node for 5, 10, 15 and 20 residue pattern lengths. Information content (IC) and frequency score were used in motif selection.

Results and Discussion

The final PF03552 codon alignment comprised 721 sequences, was 2160 nucleotides long, and contained 2155 variable sites. The final PF00535 codon alignment comprised 281 sequences, was 1830 nucleotides long and contained 1702 variable sites. Twenty two fully sequenced angiosperm species were sampled, including the monocots *Oryza sativa* (rice)*, Hordeum vulgare* (barley)*, Zea Mays* (corn)*, Brachypodium distachyon, Sorghum bicolor* (great millet)*, Setaria italica* (foxtail millet)*, Panicum virgatum* (switchgrass)*, Musa acuminata* (banana), *Phoenix dactylifera* (date palm), and the eudicots *Arabidopsis thaliana, Populus trichocarpa* (poplar)*, Glycine Max* (soya)*, Prunus persica* (peach)*, Capsella rubella* (pink shepherd's-purse), *Thellungiella halophila, Carica papaya* (papaya)*, Citrus sinensis* (orange)*, Vitis vinifera* (grape)*, Mimulus guttatus* and *Aquilegia coerulea* (Colorado Blue Columbine)

The BEAST maximum clade credibility trees for the entire PF03552 dataset produced discordant topologies across individual runs and evolutionary models. This was not surprising given the size of the dataset and highly parameter-rich models employed. We therefore focused our analyses on the unrooted ML trees for PF03552 and PF00535 and the Bayesian trees for reduced taxon sets corresponding to each subfamily. BEAST tree branching order corresponded well with unrooted ML trees except for *CslG* and *CslE* clades, in which extensive paralogous duplications made the reconstruction of relationships problematic but did not impact proposed nomenclature.

All BEAST runs yielded effective sample size (ESS) values above 200 for all important parameters (i.e. likelihood, posterior and mean branch rate). The ML trees from independent analyses showed a convergence of likelihood values with a median difference of -1.22 for PF03552 and -0.64 for PF00535. Figure 1 shows the final amino acid best known ML tree for PF03552 (Figure 1A) and PF00535 (Figure 1B) and identifies major gene lineages that existed prior to the monocot-eudicot divergence. Previous observations of extensive paralogous duplication in *CslB, CslH, CslE, CslG, CslJ* and *CslM* clades in extant species (Yin et al., 2009) are confirmed in our analyses. In addition, we resolve the putative angiosperm ancestor to contain seven *CesA,* four *CslD,* three *CslA,* three *CslC* and two *CslE* sequences with single copies of *CslB/H* and *CslG* present during much of angiosperm diversification. Precisely when the *Poaceae*-specific *CslF* family and *CesA10* (herein re-named *CesA8*) (Schwerdt et al., 2015) split from other angiosperms remains to be determined, and the eudicot-specific *CslM* family has previously been shown to have undergone substantial duplication and loss and remains partially unresolved (Schwerdt et al., in preparation). The PF03552 tree also reveals substantial gene loss in the monocots, with minimal representation in the *CslG* and *CslJ/CslM* families. Contributing to this topological asymmetry is the (1,3;1,4)-β-glucan associated *Poaceae*-specific *CslF* family that, along with *CslM*, are the only monocot or eudicot specific major PF03552 clades to have undergone expansion early in their respective evolutionary histories. The sequencing of more non-*Poaceae* monocots will provide the opportunity to explore these relationships further.

Based on these phylogenetic results, we have expanded the existing taxonomic classification (Richmond and Somerville, 2000; Hazen et al., 2002; Farrokhi et al., 2006; Schwerdt et al., in prep), by proposing a new unifying nomenclature that assigns operational taxonomic units (OTU; groups of similar sequences) according to ancestral gene duplications and resolved phylogenetic relationships.

We identified 72 OTU across the *CslA*, *CslB*, *CslC*, *CslD*, *CslE*, *CslF*, *CslG*, *CslH*, *CslJ*, *CslM*, and *CesA* families for the twenty two sampled species. Thirty of these are reciprocally monophyletic (descended from a common ancestor), the *CesA* subfamily comprises eight major lineages that pre-date the eudicot-monocot divergence. The deepest split is found between a group comprising *CesA1, CesA2, CesA3, CesA4* and another containing *CesA5, CesA6, CesA7.* This relationship corresponds to the division between the primary and secondary cell wall co-expressed genes (Burton et al., 2004; Carroll and Specht, 2011). The PF03552 ML tree (Figure 1A) supports these relationships, however it also positions the primary *CesA* clade nested within the secondary *CesA* clade, supporting previous observations that the secondary cell wall associated *CesA* genes are the ancestral lineage (Schwerdt et al., 2015). Additionally, only the primary cell wall associated *CesA* genes have undergone major duplication events within the monocots and eudicots following their divergence (*CesA4_A, CesA4_B, CesA5_A, CesA5_B, CesA5_C, CesA5_D, CesA6_A, CesA6_B, CesA6_C, CesA7_A, CesA7_B*), and thus appear to have diversified subsequent to the appearance of monocot and eudicot lineages. We also recognise *CesA8* as a *Poaceae*-specific highly diverged *CesA* family, they took significantly longer to converge and produced overall lower node posterior probability values. Figure 3 shows our amino acid BEAST tree for the *CslD* family*.* The family is resolved in our analyses to have four lineages that existed prior to the monocot-eudicot divergence, *CslD1, CslD2, CslD3* and *CslD4.* Additionally, only the *CslD1* clade has undergone any subsequent gene expansion with a major duplication in the ancestors of eudicots (*CslD1_A* and *CslD1_B)* and monocots (*CslD1_C* and *CslD1_D*).

The *CslF* family remains a *Poaceae*-specific clade, with no members recovered from the non-grass monocots *Musa acuminata* (banana) and *Phoenix dactylifera* (date palm). We have characterised the *CslF* numerically where lineages existed in the ancestor of all *Poaceae,* ranking according to branching order as determined in both the Bayesian analyses and PF03552 ML trees. As shown in Figure 4, our nucleotide BEAST tree recovered seven major gene lineages, *CslF1, CslF2, CslF3, CslF4, CslF5, CslF6* and *CslF7.* The existing *CslF6* gene family that encodes (1,3;1,4)-β-glucan synthases is the first to branch following the original *CslD-CslF* duplication. We have classified this family as *CslF1* followed by *CslF2* (originally *CslF7*) and then the *CslF* genes that comprise homologous clusters across all sampled species (Schwerdt et al., 2015).

CslE, CslG, CslM, CslJ, CslB, and CslH

Extending previous observations (Yin et al., 2009), the *CslH/CslB/CslE/CslJ/CslM* clade is notable by the presence of several independent gene duplication events after angiosperm speciation. As shown in Figure 5, our nucleotide BEAST tree of the *CslB* and *CslH* families reveals no major duplication events prior to the monocot-eudicot divergence. Indeed, excluding *Capsella rubella* and *Populus trichocarpa,* every species present in the eudicot-specific *CslB* clade has paralogous duplications that have occurred following the appearance of extant species. *CslH* is the monocot-specific sister clade to *CslB*, and is well represented with paralogues except in the Pooideae (*Brachypodium distachyon* and *Hordeum vulgare*). This suggests only very recent duplication in a gene family that has been represented by a single gene copy for most of its evolutionary history. Thus, we only retain the root symbols for this family and classify paralogues according to chromosomal order where available. Deep *CslB* nodes are poorly supported and show disagreement between ML and Bayesian analyses. However, this is unsurprising because with data constructed from exclusively post-speciation duplications, we are resolving a species tree from ancient and highly diverged single copies of genes. Importantly, this poorly supported topology does not impact our nomenclatural recommendations.

The *CslE* clade also contains this characteristic paralogous duplication structure. As shown in Figure 6, our nucleotide BEAST tree of *CslE* has two well-resolved lineages that existed prior to the eudicot and monocot divergence, *CslE1* and *CslE2.* However, the monocot copy of *CslE1* appears to have been lost, with the remaining eudicot family members comprising a single clade. *CslE2* is recovered to contain two reciprocally monophyletic eudicot and monocot clades, *CslE2_A* (eudicot)*, CslE2_B* (eudicot)*, CslE2_C* (monocot)*, CslE2_D* (monocot)*.* While support for deep *CslE2* nodes is poor, alternative topologies sampled during model testing observed only changes in branch order, particularly with the early diverging *Musa acuminata* (banana) sequences, and did not affect the proposed nomenclature.

Extensive post-speciation duplication observed in the *CslG* family, notably in the *Vitis vinifera* sequences, is consistent with previous observations (Giannuzzi, G., et al.

2011). Our BEAST *CslG* tree shows a single OTU with nomenclature addressing only paralogues (Figure 7). *CslG* has been reported as a eudicot-specific clade (Yin et al., 2009), however our analyses presented here show that the monocots *Musa acuminata* and *Panicum virgatum* also contain *CslG* sequences. As noted for the *CslB* family, ML and Bayesian analyses produce conflicting topologies but are not problematic for our nomenclatural recommendations.

Although sister to *CslG,* the *CslJ* and *CslM* clades are not compromised of extensive paralogous duplications. We have recently shown the *CslJ* to be a Poales-specific gene family that encodes a (1,3;1,4)-β-glucan synthase (Schwerdt et al., in preparation) and this is resolved in our nucleotide BEAST tree to be sister to three *CslM* eudicot specific OTU, *CslM_A, CslM_B* and *CslM_C* (Figure 8). Notably, *CslJ* sequences have been lost in *Oryza sativa* (rice) and *CslM* sequences lost in *Arabidopsis thaliana*.

CslA and CslC

There is phylogenetic evidence to suggest that *CslA* and *CslC* form an independent lineage with respect to the other *CesA* superfamily members (Nobles and Brown, 2004). However, as shown in Figure 1B, *CslA* and *CslC* are separated by a very long molecular branch, indicating substantial divergence since their common ancestor. Figure 9 shows the protein BEAST tree of the *CslA* family and reveals three major lineages that were present before the monocot-eudicot divergence, with *CslA1* and *CslA2* sister to *CslA3. CslA1* is recovered as a small clade with only five eudicot family representatives. *CslA2* comprises the largest clade within *CslA* with the eudicot

CslA2_A as sister to a clade comprising the monocot specific *CslA2_B* and eudicot specific *CslA2_C* and *CslA2_D. CslA3* is recovered as a large monocot-specific gene family with six duplications occurring before sampled extant monocots diverged, *CslA3_A, CslA3_B, CslA3_C, CslA3_D, CslA3_E* and *CslA3_F.*

As seen in Figure 10, our protein BEAST tree of the *CslC* gene family reveals three major lineages that existed prior to the monocot-eudicot divergence, *CslC1, CslC2* and *CslC3. CslC1* is a small clade with only a *Musa acuminata* sequence representing the monocots. Sisters to *CslC1* are *CslC2* and *CslC3*, both comprising major gene duplications following the monocot-eudicot split. *CslC2* has a eudicot-specific clade sister to three monocot specific families, namely *CslC2_A, CslC2_B* and *CslC2_C. CslC3* comprises two major divisions; the eudicot-specific *CslC3_D* sister to two monocot-specific clades, *CslC3_A, CslC3_B* and the eudicot-specific *CslC3_C.* We can infer that the monocot copy of *CslC3_D* was lost soon after the monocot-eudicot divergence.

To formally categorise our operational taxonomic units we conducted a discriminatory motif search to identify family specific sequence patterns. We used DEME (Redhead and Bailey, 2007) to discover protein motifs present in a positive set (each OTU) and a negative set (the entire sequence set with the OTU sequences excluded). DEME was used to search each node in the tree for 5, 10, 15 and 20 residue long motifs. The identified discriminatory motifs for our OTUs are given in Table 1.

Conclusion

In this paper we performed phylogenetic analyses on cellulose synthase family genes from a broad sample of angiosperm species and used resolved relationships to propose a new nomenclature for this important gene family. Specifically, we have identified clades that existed prior to monocot-eudicot divergence and thus represent gene families present at the origin of the angiosperm radiation. We have retained the existing root symbols present in the literature but have extended these classifications: firstly by numbering lineages that predate the monocot-eudicot divergence, and secondly by using character symbols for major eudicot or monocot duplications. The new nomenclature, and the previous *Arabidopsis thaliana* and *Oryza sativa* designations, are presented in Table 1.

In this nomenclatural system, we have assigned sequential numbers to related gene families within the previously named *CesA* and *Csl* groupings. This limits extensibility if sequential order is to be maintained for newly discovered genes that are recovered outside of numbered groups in the phylogeny. However, because we use additional character identifiers for subsequent major duplications within monocots or eudicots, any attempt to preserve numerical order will likely cause confusion. Thus, we propose that new genes resulting from ancestral duplications prior to the eudicot-monocot divergence be assigned the next number available in the sequence regardless of their phylogenetic position with respect to other numbered groups.

References

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol **10**: 1–6

Burton R a, Shirley NJ, King BJ, Harvey AJ, Fincher GB (2004) The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. Plant Physiol **134**: 224–236

Burton R a, Wilson SM, Hrmova M, Harvey AJ, Shirley NJ, Medhurst A, Stone BA, Newbigin EJ, Bacic A, Fincher GB (2006) Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1,3;1,4)-β-D-glucans. Science **311**: 1940–2

Camacho C, Coulouris G, Avagyan V, MA N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics **10**: 421

Carroll A, Specht CD (2011) Understanding Plant Cellulose Synthases through a Comprehensive Investigation of the Cellulose Synthase Family Sequences. Front Plant Sci **2**: 5

Cocuron J-C, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, Wilkerson CG (2007) A gene from the cellulose synthase-like C family encodes a β-1,4 glucan synthase. Proc Natl Acad Sci U S A **104**: 8550–8555

Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol **10**: 210

Darriba D, Taboada GL, Posada D (2011) Supplementary material for “ProtTest 3 : fast selection of best-fit models of protein evolution” Summary : Availability : 1–4

Doblin MS, Pettolino FA, Wilson SM, Campbell R, Burton RA, Fincher GB, Newbigin E, Bacic A (2009) A barley cellulose synthase-like CSLH gene in transgenic Arabidopsis. 106:

Drummond A, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **7**: 214

Dwivany FM, Yulia D, Burton RA, Shirley NJ, Wilson SM, Fincher GB, Bacic A, Newbigin E, Doblin MS (2009) The CELLULOSE-SYNTHASE LIKE C (CSLC) family of barley includes members that are integral membrane proteins targeted to the plasma membrane. Mol Plant **2**: 1025–1039

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**: 1792–1797

Farrokhi N, Burton RA, Brownfield L, Hrmova M, Wilson SM, Bacic A, Fincher GB (2006) Plant cell wall biosynthesis: Genetic, biochemical and functional genomics approaches to the identification of key genes. Plant Biotechnol J **4**: 145–167

Felsenstein J (1993) {PHYLIP}: phylogenetic inference package, version 3.5c.

Fincher GB (2009) Revolutionary times in our understanding of cell wall biosynthesis and remodeling in the grasses. Plant Physiol **149**: 27–37

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res **39**: W29–37

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res **40**: 1178–1186

Hazen SP, Scott-craig JS, Walton JD (2002) Cellulose synthase-like (CSL) genes of rice Cellulose Synthase-Like Genes of Rice 1. doi: 10.1104/pp.010875.synthesize

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Res **12**: 1611–1618

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: Genome-scale protein function classification. Bioinformatics **30**: 1236–1240

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol **29**: 1695–1701

Lee EK, Cibrian-Jaramillo A, Kolokotronis S-O, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, McCombie WR, Martienssen GC, DeSalle R (2011) A functional phylogenomic view of the seed plants. PLoS Genet **7**: e1002411

Liebert MA, Chen K, Durand D, Farach-colton M, Al CET (2000) NOTUNG : A Program for Dating Gene Duplications. **7**: 429–447

Liepman AH, Wilkerson CG, Keegstra K (2005) Expression of cellulose synthase-like (Csl) genes in insect cells reveals that CslA family members encode mannan synthases. Proc Natl Acad Sci U S A **102**: 2221–6

Mayer KFX, Waugh R, Brown JWS, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature **491**: 711–6

Nobles DR, Brown RM (2004) The pivotal role of cyanobacteria in the evolution of cellulose synthases and cellulose synthase-like proteins. Cellulose **11**: 437–448

Posada D (2008) jModelTest: Phylogenetic model averaging. Mol Biol Evol **25**: 1253–1256

Redhead E, Bailey TL (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinformatics **8**: 385

Richmond T a, Somerville CR (2000) The cellulose synthase superfamily. Plant Physiol **124**: 495–8

Scheible WR, Pauly M (2004) Glycosyltransferases and cell wall biosynthesis: Novel players and insights. Curr Opin Plant Biol **7**: 285–295

Schwerdt JG, MacKenzie K, Wright F, Oehme D, Wagner JM, Harvey AJ, Shirley NJ, Burton R a, Schreiber M, Halpin C, Zimmer J, Marshall DF, Waugh R, Fincher GB (2015) Evolutionary Dynamics of the Cellulose Synthase Gene Superfamily in Grasses. Plant Physiol **168**: 968–983

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol **7**: 539

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**: 1312–1313

Wheelan SJ, Church DM, Ostell JM (2001) Spidey: A tool for mRNA-to-genomic alignments. Genome Res **11**: 1952–1957

Whelan S, Allen JE, Blackburne BP, Talavera D (2014) ModelOMatic: Fast and Automated Model Selection between RY, Nucleotide, Amino Acid, and Codon Substitution Models. Syst Biol **64**: 42–55

Yin Y, Huang J, Xu Y (2009) The cellulose synthase superfamily in fully sequenced plants and algae. BMC Plant Biol **9**: 99

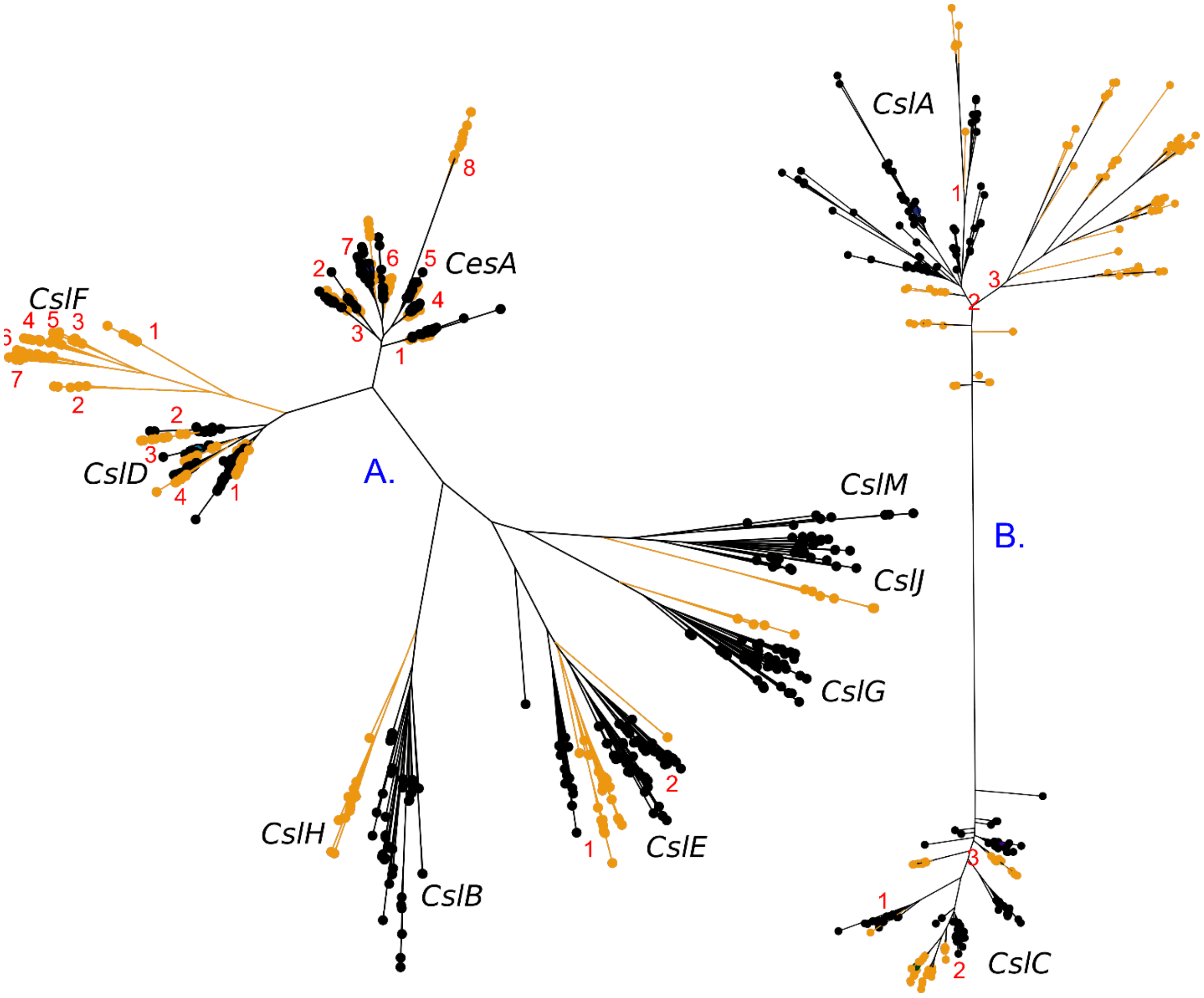


Figure 1: RAxML best known maximum likelihood tree (amino acid) showing major lineages and existing nomenclature root symbols for the *CesA* superfamily group A (PF03552) and group B (PF00535) from 22 fully sequenced angiosperms. Eudicot taxa are coloured black, monocot taxa are coloured orange. Lineages that have existed prior to the monocot-eudicot divergence are numbered in red. Additionally clades that existed prior to appearance of extant species in the *Poaceae*-specific *CslF* family are numbered in red.

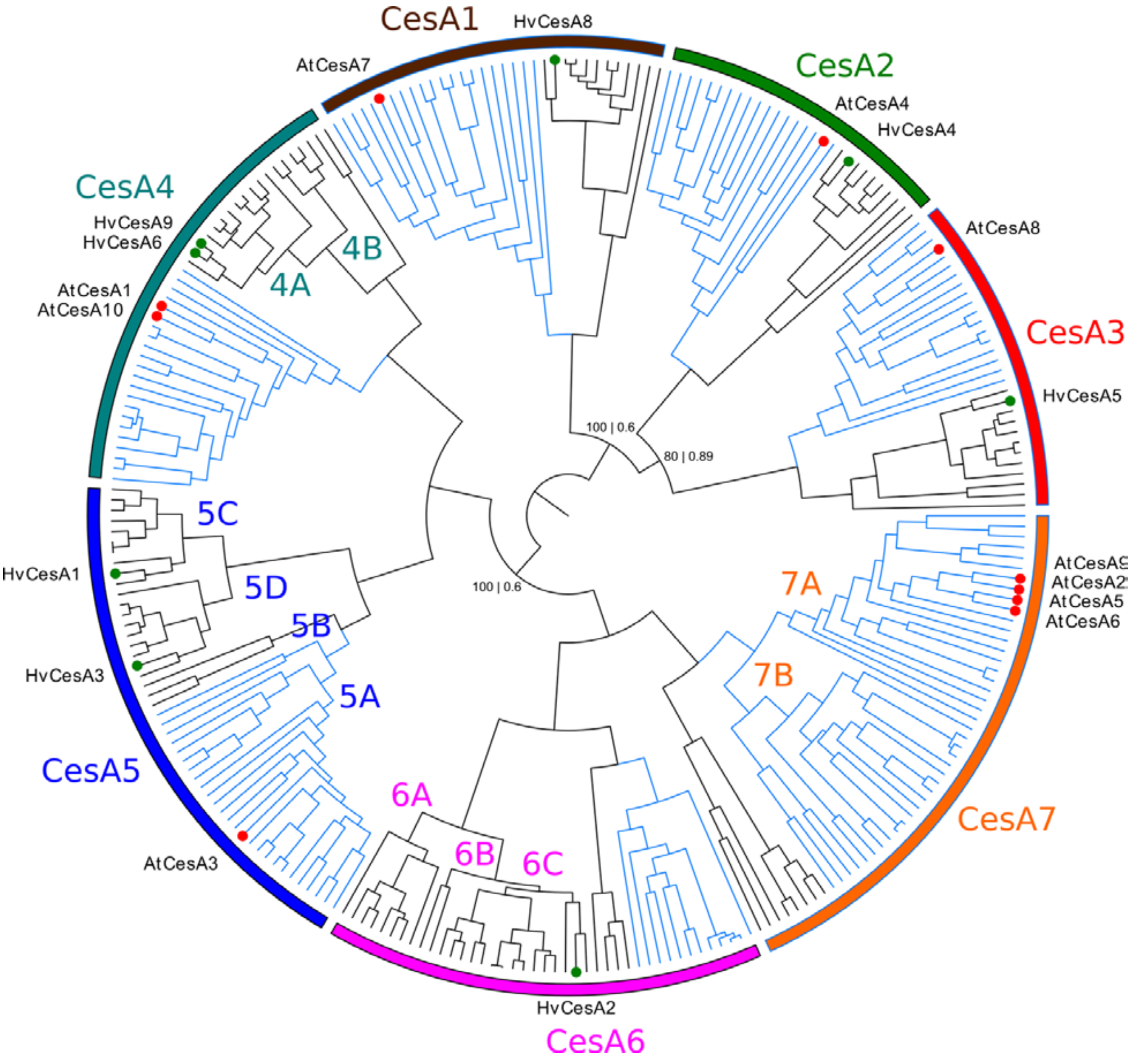


Figure 2: BEAST maximum clade credibility tree for the *CesA* family from 22 fully sequenced angiosperms. Branches for eudicot and monocot genes are coloured blue and black respectively. Lineages that existed prior to the monocot-eudicot divergence are identified numerically. Subsequent entire monocot or eudicot duplications are identified by a single character suffix (A-D). Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RAxML maximum likelihood support values indicated as bootstrap percentage. Unmarked deep nodes have a

posterior probability of >0.95 and bootstrap support of >90%. Existing barley and arabidopsis gene labels are indicated at tips.

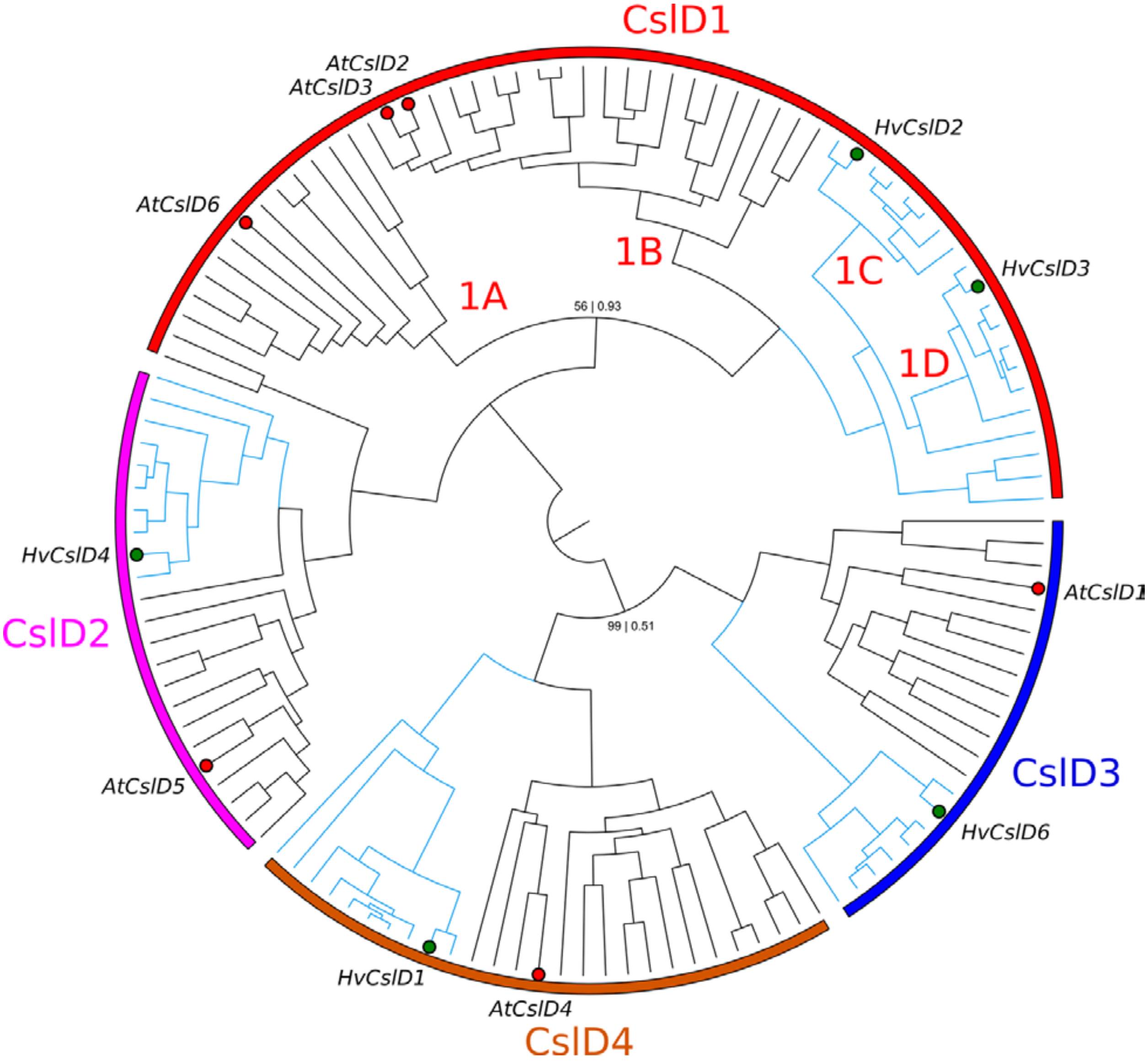


Figure 3: BEAST maximum clade credibility tree for the *CslD* family from 22 fully sequenced angiosperms. Branches for eudicot and monocot genes are coloured black and blue respectively. Lineages that existed prior to the monocot-eudicot divergence are identified numerically. Subsequent entire monocot or eudicot duplications are

identified by a single character suffix (A-D). Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RaxML maximum likelihood support values indicated as bootstrap percentage. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap support of >90%. Existing barley and arabidopsis gene labels are indicated at tips.

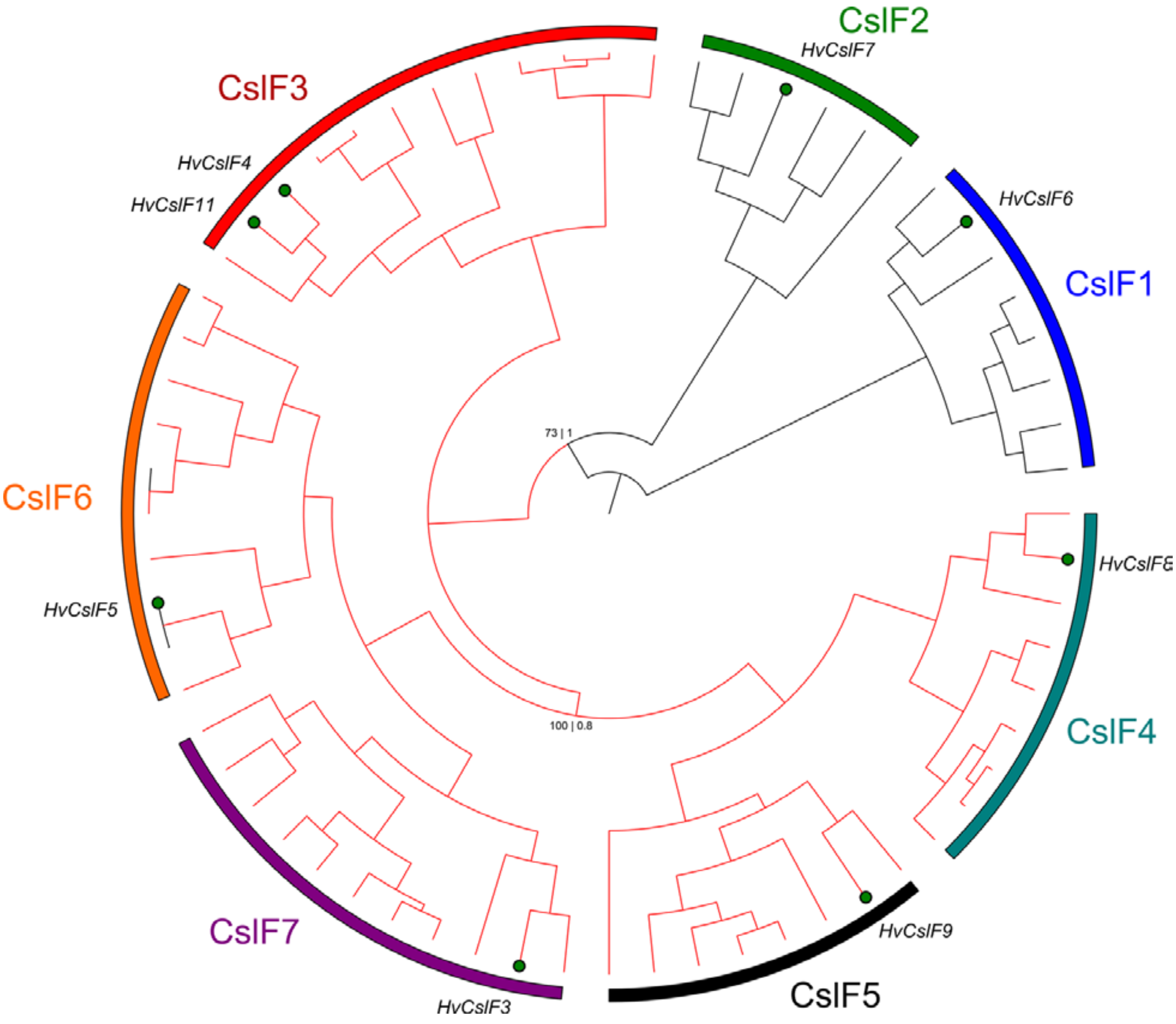


Figure 4: BEAST maximum clade credibility tree for the *CslF* family from seven fully sequenced *Poaceae*. Branches for the genes located in the conserved *CslF* are coloured red*.* Posterior probability support values (0.1 - 1) are presented for deep

nodes with the corresponding RaxML maximum likelihood support values indicated as bootstrap percentage. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap support of >90%. Existing barley gene labels are indicated at tips.

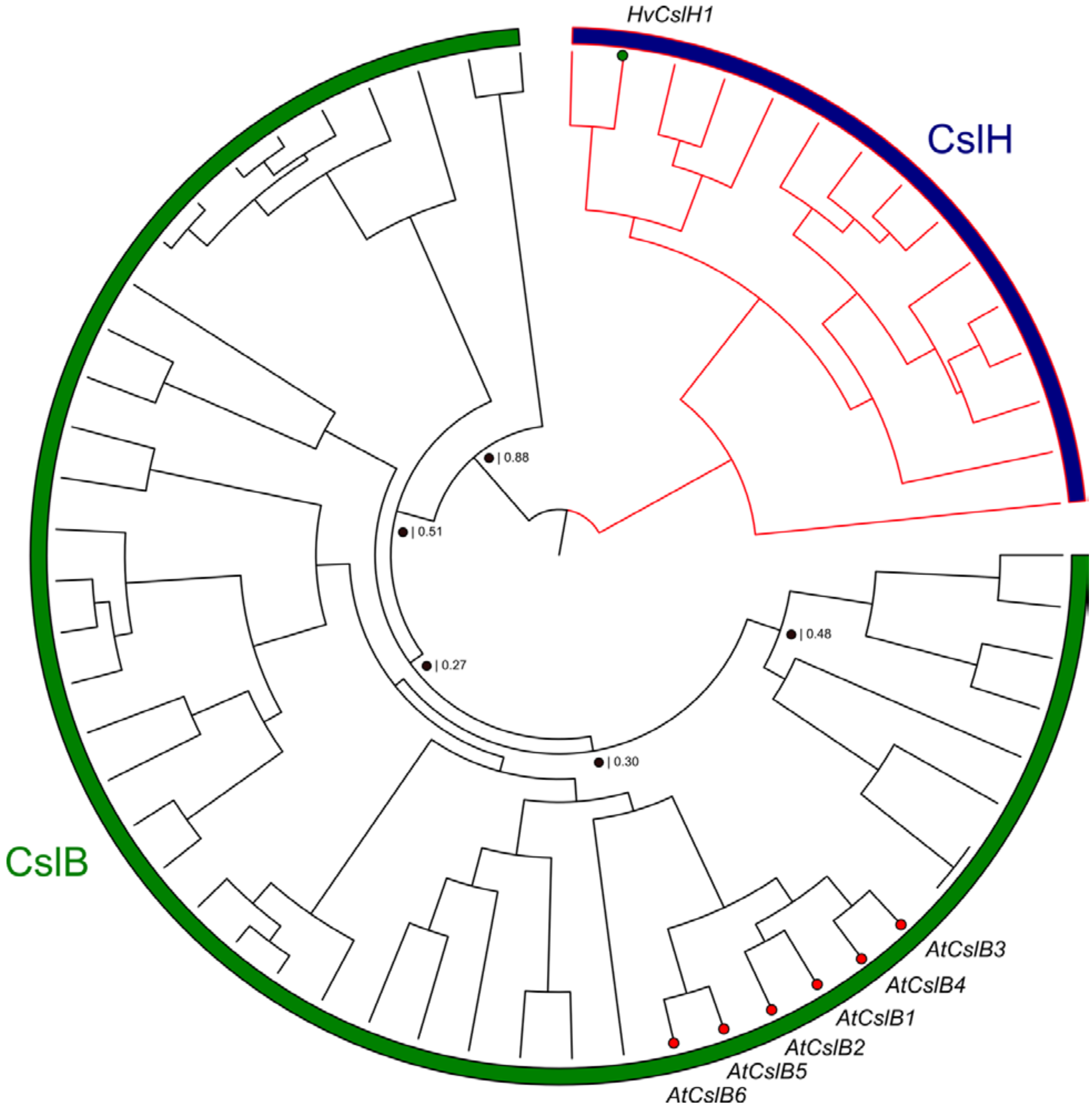


Figure 5: BEAST maximum clade credibility tree for the *CslB* and *CslH* families from 22 fully sequenced angiosperms. Branches for eudicot and monocot genes are coloured

black and blue respectively. Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RAxML maximum likelihood support values indicated as bootstrap percentage. A black circle indicates topology disagreement between Bayesian and maximum likelihood analyses. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap support of >90%. Existing barley and arabidopsis gene labels are indicated at tips.

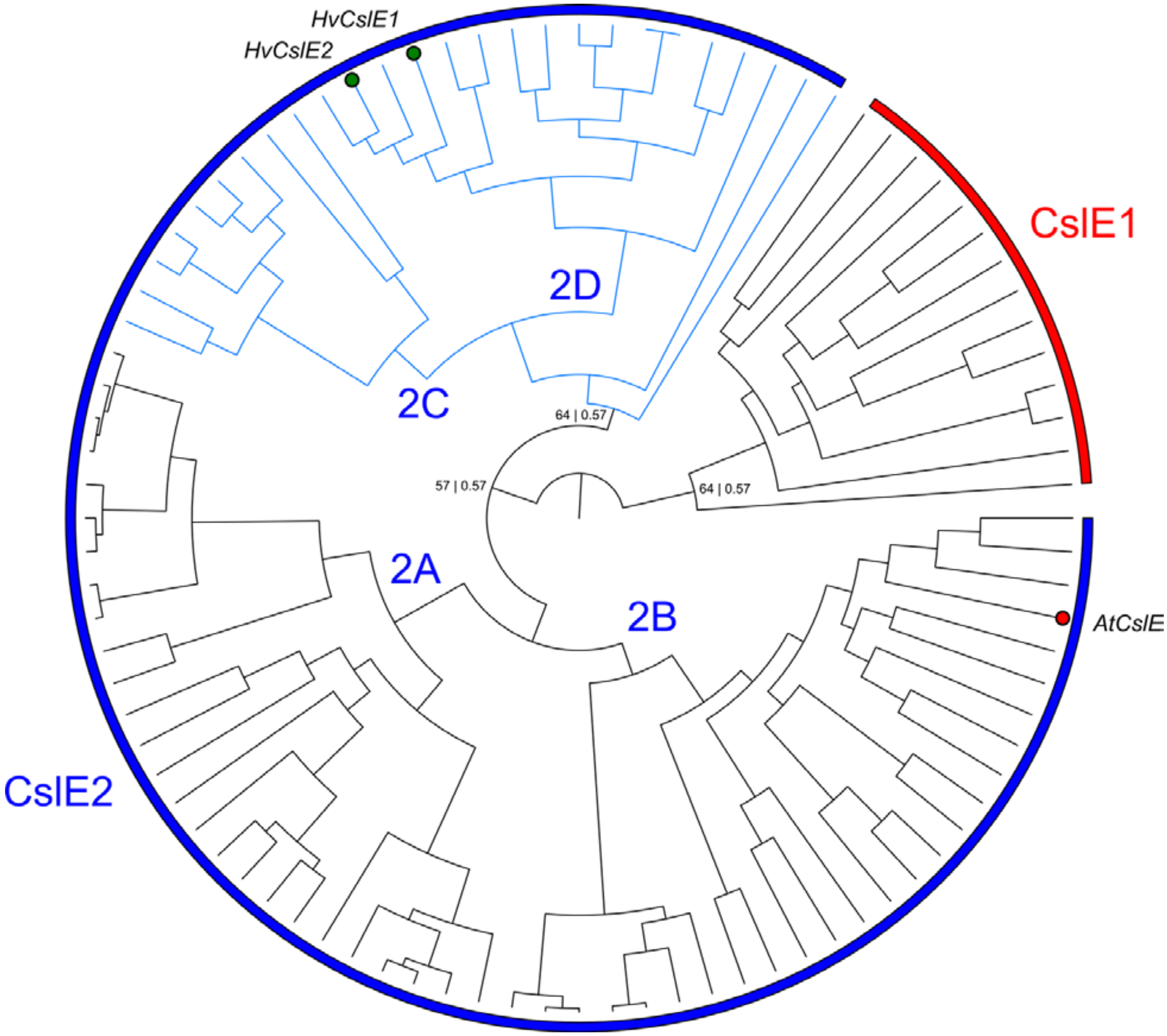


Figure 6: BEAST maximum clade credibility tree for the *CslE* family from 22 fully sequenced angiosperms. Branches for eudicot and monocot genes are coloured black

and blue respectively. Lineages that existed prior to the monocot-eudicot divergence are identified numerically. Subsequent entire monocot or eudicot duplications are identified by a single character suffix (A-D). Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RaxML maximum likelihood support values indicated as bootstrap percentage. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap support of >90%. Existing barley and arabidopsis gene labels are indicated at tips.

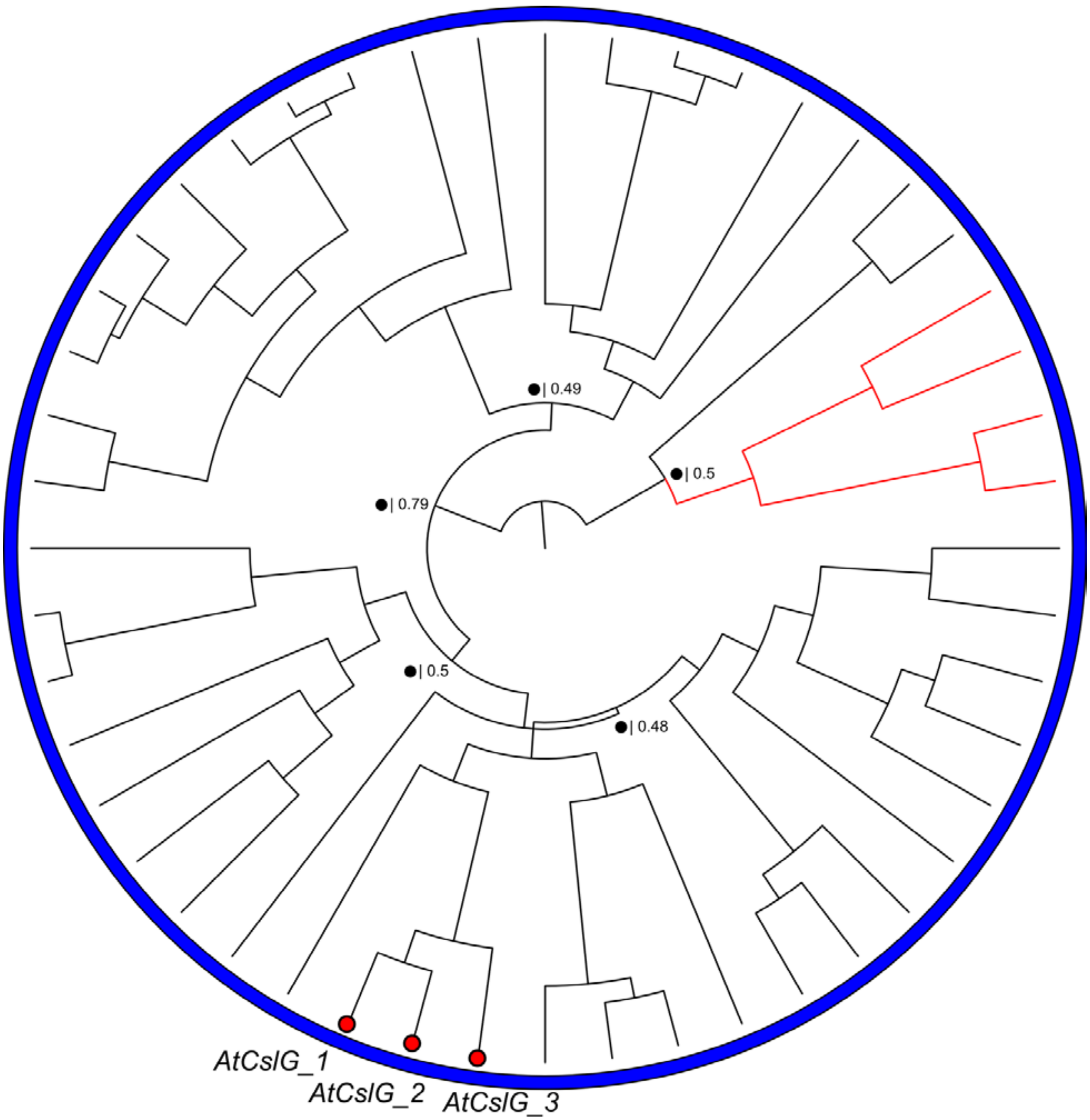


Figure 7: BEAST maximum clade credibility tree for the *CslG* family. Branches for eudicot and monocot genes are coloured black and red respectively. Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RaxML maximum likelihood support values indicated as bootstrap percentage. A black circle indicates topology disagreement between Bayesian and maximum likelihood analyses. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap

support of >90%. Existing arabidopsis gene labels are indicated at tips.

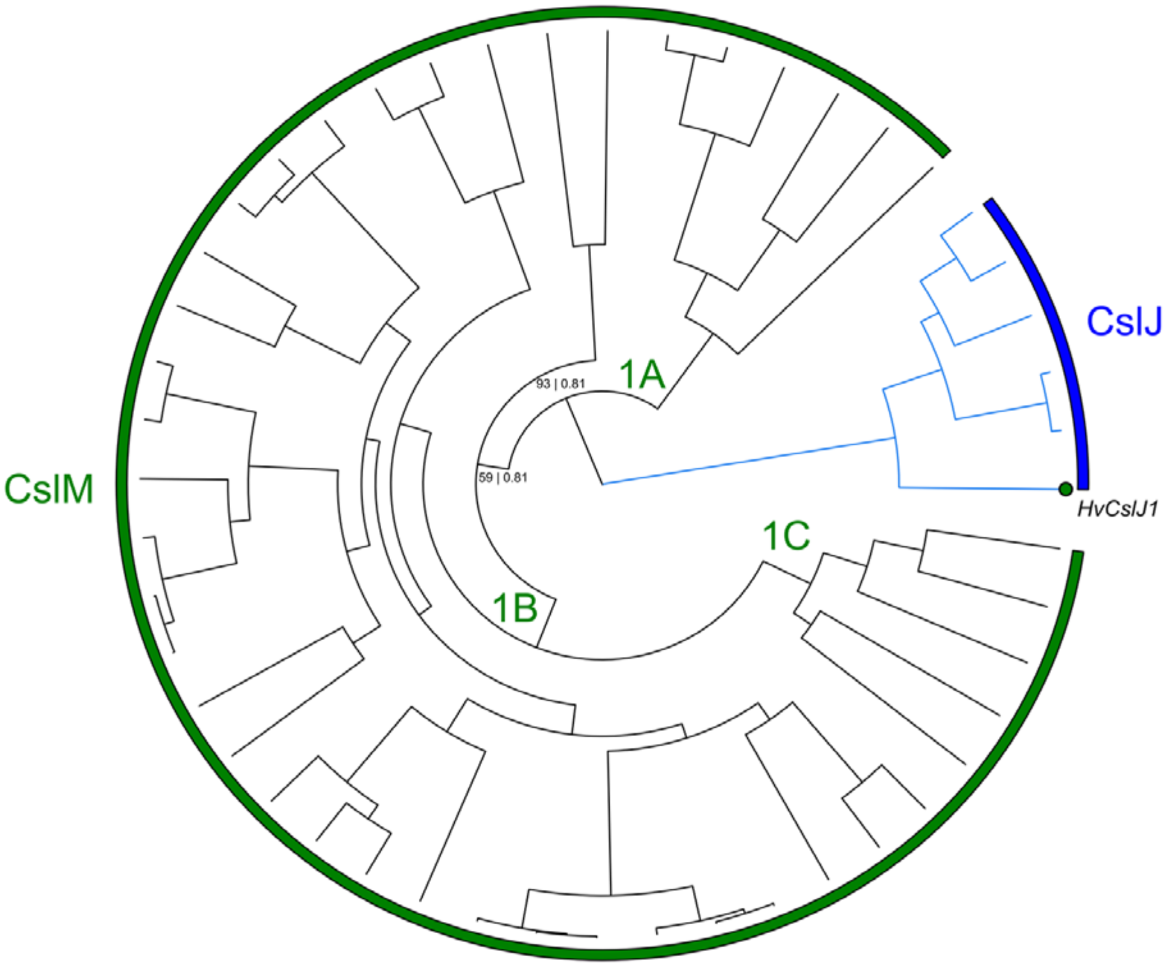


Figure 8: BEAST maximum clade credibility tree for the *CslM* and *CslJ* families from 22 fully sequenced angiosperms. Branches for eudicot and monocot genes are coloured black and blue respectively. Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RAxML maximum likelihood support values indicated as bootstrap percentage. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap support of >90%. Existing barley gene labels are indicated at tips.

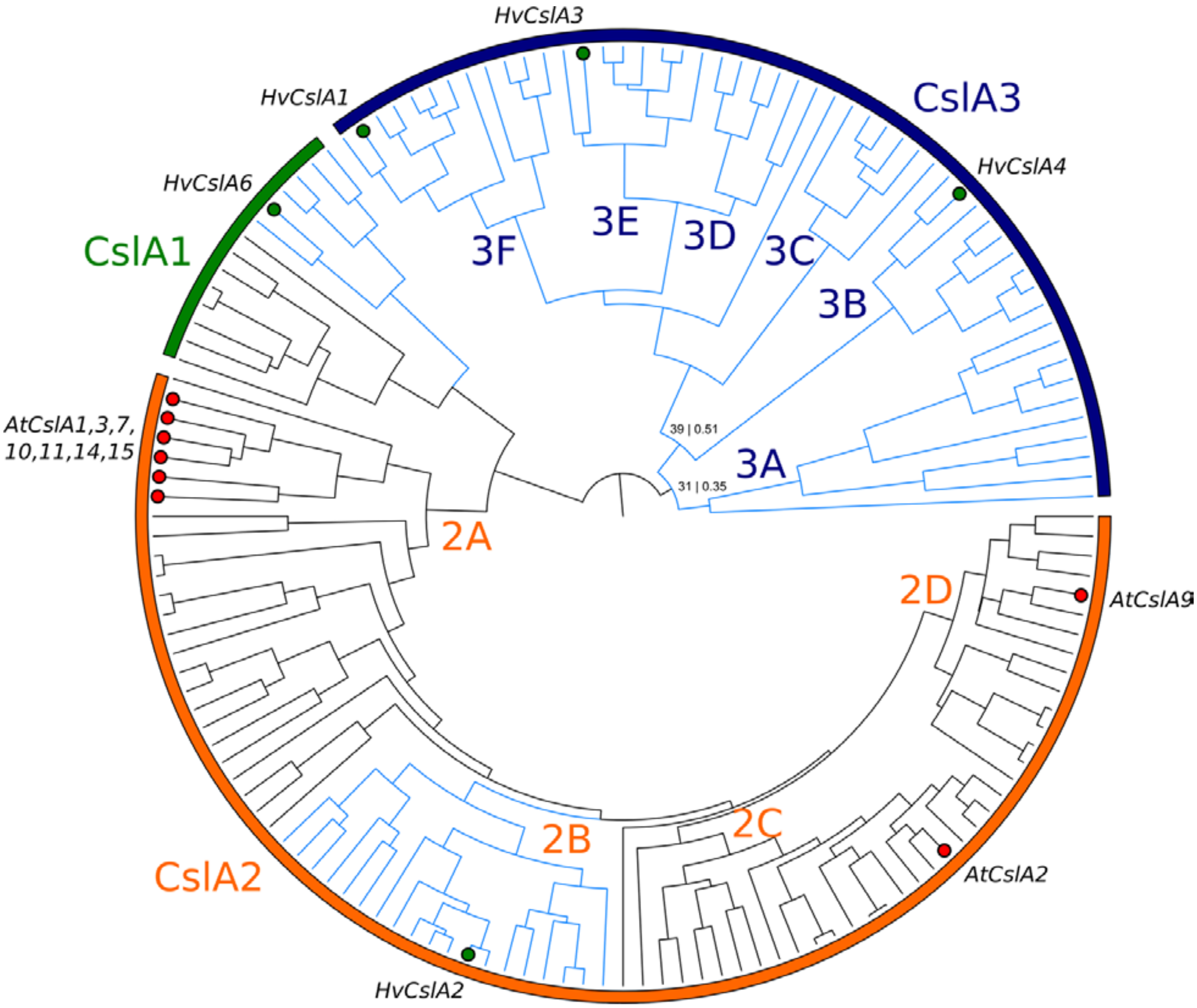


Figure 9: BEAST maximum clade credibility tree for the *CslA* family from 22 fully sequenced angiosperms. Branches for eudicot and monocot genes are coloured black and blue respectively. Lineages that existed prior to the monocot-eudicot divergence are identified numerically. Subsequent entire monocot or eudicot duplications are identified by a single character suffix (A-D).Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RaxML maximum likelihood support values indicated as bootstrap percentage. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap support of >90%. Existing barley gene labels are indicated at tips.

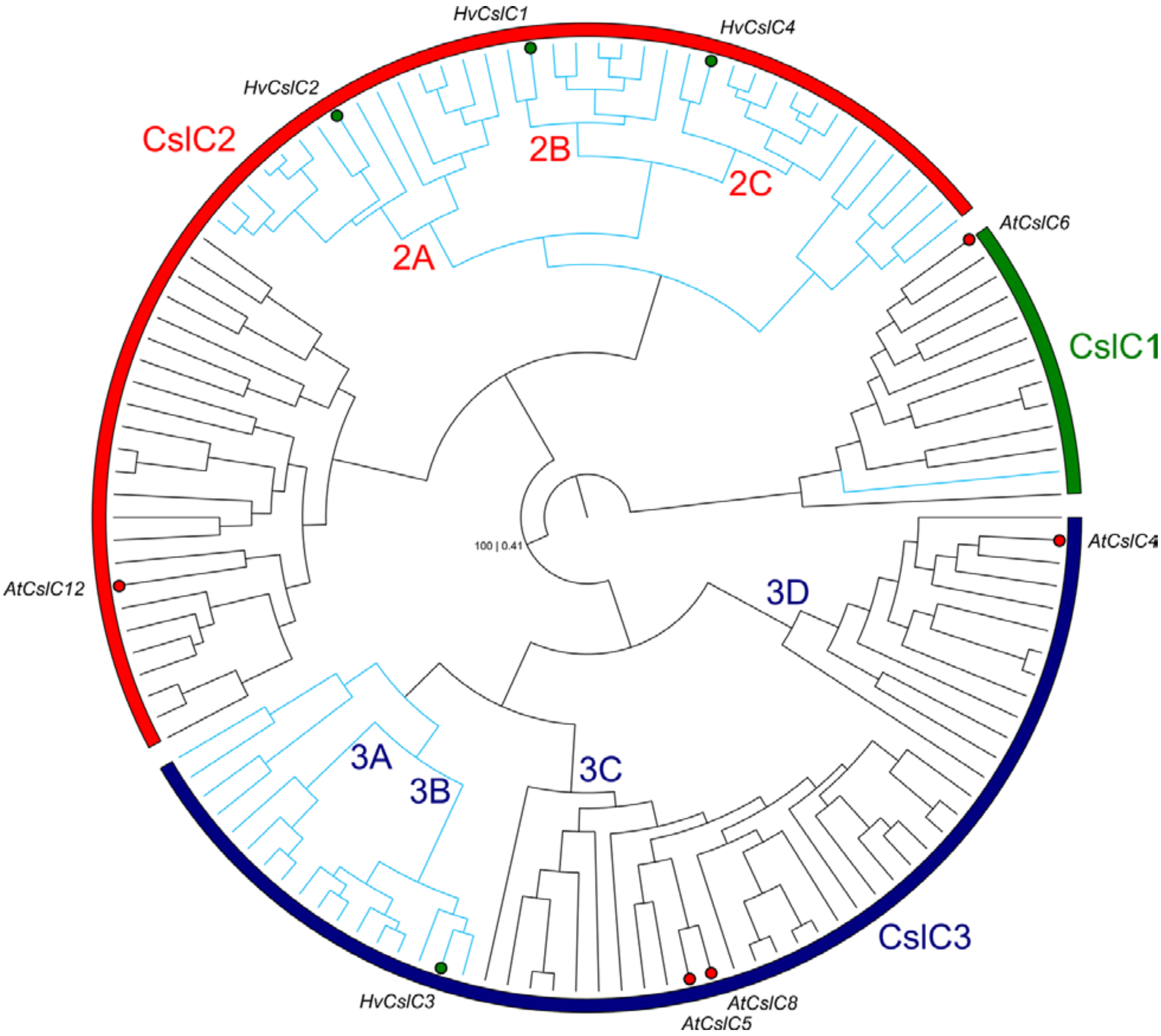


Figure 10: BEAST maximum clade credibility tree for the *CslC* family from 22 fully sequenced angiosperms. Branches for eudicot and monocot genes are coloured black and blue respectively. Lineages that existed prior to the monocot-eudicot divergence are identified numerically. Subsequent entire monocot or eudicot duplications are identified by a single character suffix (A-D).Posterior probability support values (0.1 - 1) are presented for deep nodes with the corresponding RaxML maximum likelihood support values indicated as bootstrap percentage. Unmarked deep nodes have a posterior probability of >0.95 and bootstrap support of >90%. Existing barley gene

labels are indicated at tips.

Table 1. New nomenclature for existing cellulose synthase and the closely related cellulose synthase-like designations and discriminatory motifs. Existing *Arabidopsis,* rice and barley gene symbols included for reference.

OTU	Arabidopsis	Rice	Barley	Motif (5)	Motif (10)	Motif (15)	Motif (20)
CesA1	AtCesA7	OsCesA9	HvCesA8	CPCFG	LVHIHGHETH	LVLIMGHENHKPVRA	ELVLIRGHQDHKPVKALAGQ
CesA2	AtCesA4	OsCesA7	HvCesA4	LKPCG	GFEGFEGLER	EKRGLVNKDQGPDDD	ECHSRNGFGYEDLERSSLNS
CesA3	AtCesA8	OsCesA4	HvCesA5	PIWKD	LGELESYDDH	EKDLSELYRDAKREE	QITKDLAEVYRDAKREDLNS
CesA4	AtCesA1/10	OsCesA1	HvCesA6/9	KSCCG	IIVKSCCGRR	EDLEPNIIVKSCCGS	IMTQEDLEPNIIVKSCCGKR
CesA4_A	-	OsCesA1	HvCesA6/9	PPRPC	HKYPEPRGAA	GDAPAPGKPGKGAGG	HHDVKAPTPTKPGKSVNGQV

CesA4_B	-	-	-	LRPCF	FAADGKGNIE	IIQVCNNYCCENNNN	RNMVAGSCNGVIMVRYNNDG
CesA5	AtCesA3	OsCesA2/8	HvCesA1/3	IFLHY	KKPERMVSWH	PEKNKKPGFFSSLCG	PPIKVKHKKPSLLSKLCGGS
CesA5_A	AtCesA3	-	-	HKPKH	NYDKEVSHNH	YDKEVSFNHIPYLTS	APNYDKEVSHNHIPLLTSRR
CesA5_B	-	-	-	LGWHM	QRENIGPPKF	INNNQNKNNFFNIWN	VHPLSYRNPNNSRQFGNVAW
CesA5_C	-	OsCesA8	HvCesA1	GQYMI	VGKRASFPYV	TGNVGKRASFPYVNH	ITPTGHVGKKASFPYIHHAP
CesA5_D	-	OsCesA2	HvCesA3	MFTWR	LGYIPTFTHG	YILRLAHVQTTGEML	PEKMLTWRTNSGAGDDAGLT
CesA6	-	OsCesA3/5/6	HvCesA2	LLENC	LEYGGTLRCD	QHALVPSYMAQVGGH	VVVQPFFAISNVPLLTNGQM
CesA6_A	-	OsCesA6	-	KWCLS	LEGKFGLHGG	MGGGGGGARRAEAPC	FGLQGGEGHEDDPHYVAQSM

CesA6_B	-	-	-	WSDKH	KFEKLKKLFK	FEKVKRLFQTKENQA	NPNMQSWWHNVEGNELARLI
CesA6_C	-	-	-	NWCWC	PNGAQQPFQL	FNWDGNESQYGAESL	EFNWDGHESQYGADSLHGHM
CesA7	AtCesA2/5/6/9	-	-	LEVCG	KHKEASKQIH	QKCQEASKQIHALEN	NGKCKEASNQVHCLENIGRG
CesA7_A	AtCesA2/5/6/9	-	-	WCCLW	PFVTKDGPIL	VQKMNTTQMKLEKKF	ENNTIQKMNTGQMKLEKKFG
CesA7_B	-	-	-	GCGLK	IEGIDNEKSS	NAPKEMDTAAVNTEI	IDGSGFRTPSDLDPASVNPE
CslD1	AtCslD2/3/6	OsCslD1/2	HvCslD2/3	PFKTK	KDPFKNKIKH	PPAQNSQLGGSFQLW	AHRMMDSDDQEMNPALSPKK
CslD1_A	AtCslD6	-	-	GTWTI	PEVFHLKKWK	RGTGGDSSALDLAEV	TWMADSDTTCWPGTWTGSGA
CslD1_B	AtCslD2/3	-	-	CDFKL	NRGLRRGDED	DIDDEDMNPSLAQKV	IWPKDGGFGGEDDVVAPTE

CslD1_C	-	OsCslD2	HvCslD2	EHSGC	GLGGADGQPA	DPLYGSTGDEGRPLD	KETELDDIVGAGNGGADGKP
CslD1_D	-	OsCslD1	HvCslD3	YHVHI	NEEKRPVDFT	GDGGDGQPTELMTKP	AYKTTEWEDLVGLAAFTSTR
CslD2	AtCslD5	OsCslD4	HvCslD4	HHGWF	TQHNGWFGTK	HHGWFASKRMKFLLT	HHGWFGRKRVKPLQDVKTFA
CslD3	AtCslD1	OsCslD5	HvCslD6	WLKRE	RTIEYRGVFG	PLLEHPDHDTPQKFG	TPPLTGHPDHDAPQKFGKSK
CslD4	AtCslD4	OsCslD3	HvCslD1	CRFKI	DAQKDTCLCP	YLSLSREDIDMSGEL	YVSLSRDDIDMSGELSGDYA
CslF1	-	OsCslF6	HvCslF6	WTHWL	LDGEWTHWLK	CCGFPVCACAGSAAV	GCACGGFPVCACSGAAAVAS
CslF2	-	OsCslF7	OsCslF7	FFNCT	RADYKGRAWP	FFKWRVSTALVMMNS	LVLFFKWRISTALAMMSSPD
CslF3	-	OsCslF1/2/4	HvCslF4/11	WCRSD	GQATAWGLFT	TWGFFTHQSWHAVLG	GKAASWGPLTEPGWLAVLTM

CslF4	-	OsCslF8	HvCslF8	FYILF	LRGGSVSFKF	LIAIQATSTGNEKYH	RTGGARFHGGHSAGASFPTA
CslF5	-	OsCslF9	HvCslF9/12	WWITV	WSLAQVAGAA	WGWSPAQVAGAAGGL	VAALAAIASGYVAVLGVLAP
CslF6	-	-	HvCslF10	WNNWW	LKSVLAALKQ	NSVLAALKQEEGISL	LSGMLYRGRSHKEFMSDYKH
CslF7	-	OsCslF3	HvCslF3	WMDYW	MGVWTAAKKM	WAGADKAERRAAKEC	EHEAPPQGGRASQEFKNDYK
CslH	-	OsCslH1/2/3	HvCslH1	FELIY	LLLGFDDEVH	RTAWKLADLAVLSLL	VPLARMAWKLAGLAVLSLLL
CslB	AtCslB1/2/3/4/5/6	-	-	PYRYF	RQCMSYFWLF	LFAKIRFRQPLSYTW	LFGAFFAKLQFRQRMAYFWL
CslE1	-	-	-	AWWNL	KEWDCKVTKQ	TDYRTVEELEEASKV	AQEKKDYRTVEEVNTASKVI
CslE2	AtCslE	OsCslE1/2/6	HvCslE1/2	VKYGC	EALCGCKYTK	ETLCGCKYTQNYKED	CKHGRDALCGNKFDQNCEMD

CslE2_A	-	-	-	CKVLA	ARVSEEVCKV	IDKASDVEEQCKVLA	WEHHRQVREKASVLEEECKV
CslE2_B	AtCslE	-	-	YCCWG	TQWTSYMSQK	AWYAVGRISLGHVMG	AWHAVGRISPGLTMCYLTYC
CslE2_C	-	OsCslE2	-	CGYWA	VHACLDSWGG	VTAEVHPCFDSWGGM	KTCNALDHACLDSWGGMKNA
CslE2_D	-	OsCslE1/6	HvCslE1/2	YRQNW	KEDWDQGMKE	YKEDWDQGMKTQHRL	YKEDWDRGIKTEHQLQQDNK
CslG	AtCslG-1/2/3	-	-	HYACW	ATEGFLDQEF	LVLGIARAATTEGFF	GFLDEQFAQLFLKMATHANF
CslJ	-	-	HvCslJ1	QCPWP	GYRRFLCRGW	YFTGYRRFLSKGWTT	VVEDYFTGYRRFYCRGWASA
CslM	-	-	-	FHQDP	YCNDPTSARQ	AMCKHLHEVISTGGS	VSSQSQHLNEVLSTGTSWRT
CslM_A	-	-	-	CFLGC	DGMVQLMKWL	LMKWASELVQLGLSK	YFTLTHFYAVACFLYGIVPQ

CslM_B	-	-	-	FHVDK	LYCLPLWCQA	SMCYAQLALFPNYYF	HSFVHCMCYAELAHFPTYFF
CslM_C	-	-	-	WFMIF	MCFNHLALQP	YGCLALQPAYSFPLW	TPDVNSTIVTRSYTFFHYVA
CslA1	-	-	HvCslA6	HCTIY	MGMYLLHCAI	PACILVGDVRLPKPI	NKNSKIHPLEILMGMCMLHC
CslA2	AtCslA2/9	OsCslA1/9	HvCslA2	PNWGA	VQVPKWQTVV	EVQVPKWSTVYAPFI	FPQIGLVMQTRSVFIVPMLK
CslA2_A	AtCslA2/9	OsCslA1/9	HvCslA2	PNWGA	VQVPKWQTVV	EVQVPKWSTVYAPFI	FPQIGLVMQTRSVFIVPMLK
CslA2_B	-	OsCslA1/9	HvCslA2	PLRLA	MVWQGCAPVV	TILPGGPATWQVCKP	LTILPNDPETWQVRSPVMSA
CsA2_C	AtCslA2	-	-	QWKKP	RILTHELGFA	FKFANRINITELGFA	STARERFKITDRILTQELGF
CslA2_D	AtCslA9	-	-	KMFSR	IVISLIRLSG	TPKLPRFRFGDSIFV	SPLPVFAILCAIRAPLLVPL
CslA3	-	OsCslA2/3/4/5/6/7/11	HvCslA1/3/4	DFLMK	RHPSNIHIIP	TQKVGNHNKDQPLTE	TAITIETSLRHPSNIHIPPI

CslA3_A	-	OsCslA6	-	YIYRD	FIARRIISTN	VFLLISACDDCIHRT	LAEKVSMGLTSLYAKVFRRK
CslA3_B	-			YHCSE	TRFWDKYHCS	LPLVRKLRARLWEGY	LPLILRVRTRFWDGYHCSEA
CslA3_C	-	OsCslA3/4	-	GIGFC	TGFCGTTSSN	SNCKPQILEKPPCRC	SCKPQILEPPVPRCWDRCTK
CslA3_D	-	OsCslA2	-	GDHVR	TRLLETMYVD	LNVCYMNVNNFVVNL	VWGMFVIWTWVWIMTWWWNW
CslA3_E	-	-	HvCslA3	ARTPC	YITEILLALY	LAEAAWMGLASLAAR	GLASLVARLLRLRRGYGYRW
CslA3_F	-	OsCslA5/7	HvCslA1	CRQEF	APTVACILYN	RVVAPTVACVLFNVI	NGFDEPLPTAKRKGLRNRVN
CslC1	AtCslC6	-	-	KCYCD	ELNKLETTKK	SGFDELNKLEVTKKT	GFDQLNKLEVTKKAGRKTKL
CslC2	AtCslC12	OsCslC1/7/9/10	HvCslC1/2/4	DCRCE	DEKQAKHNRI	GDLIALKEKKQKHNR	ALLPPKEHKQQRGASATNIE

CslC2_A	-	OsCslC9/10	HvCslA2	DIIKC	LTPPKELRHH	LLAAAPPKELRKHKT	KILAASVDNFHGSWVRFHAT
CslC2_B	-	OsCslC7	HvCslC1	LVEKH	LGLVEKHSVQ	AAAAAAGVMRTRLDY	AMDVNAAAAVAGIMRTKLDY
CslC2_C	-	OsCslC1	HvCslC4	GYASW	NAAGKGSDDV	APNVPDLAAFEELYA	WHLTAAAPDLESVEGIYASA
CslC3	AtCslC5/8	OsCslC2/3	HvCslC3	LSKFC	HWVYMAWMLF	NAYSLVEVNDPESTM	PNYSLVEIDGPGSASENQEK
CslC3_A	-	OsCslC2	-	GYLHP	LGQGRGWLLY	DKKCFTLPQLQKQLP	DRKCISLPQLQNQLPEKEEL
CslC3_B	-	OsCslC3	HvCslC3	HLRHL	DLPKHLRHLR	HLPGHLRHLPHNLRE	DHPKHLQHHRHHLPENLRHP
CslC3_C	AtCslC5/8	-	-	WHYFN	HYQDNPNLHI	WHYQDNPNLHIPHAT	VTMESEKLGKPKVLNSAIRD
CslC3_D	AtCslC4	-	-	QWCYI	FPEKQKAASP	DSLLFPQKQKAVSPK	KISGVDSNLFLEKQKAAATK

Supplementary table 1. *CesA* and *Csl* homologues for each identified OTU.

OTU	**Loci**
CesA1	Glyma.06G225500.1, Pavir.Bb02205.1, Potri.018G103900.1, Solyc07g005840.2.1, 29482.m000171, Migut.J01374.1, hv_contig_54778_CesA8, Glyma.17G072200.1, Gorai.009G009700.1, Si028761m, AT5G17420.1, Potri.006G181900.1, GRMZM2G002523_T01, GRMZM2G011651_T01, Glyma.02G205800.1, Glyma.06G225400.1, GSVIVT01023643001, Glyma.04G142700.1, orange1.1g039060m, GSMUA_Achr7G19410_001, GRMZM2G142898_T01, LOC_Os09g25490.1, orange1.1g039678m, Glyma.04G153700.1, ppa000618m, Pavir.J12997.1, Sobic.002G205500.1, Aquca_048_00031.1, Gorai.001G044700.1, Cucsa.178930.1, Bradi4g30540.1, Eucgr.C00246.1, PDK_30s779251g002
CesA2	Bradi3g28350.1, ppa000641m, Pavir.J21370.1, Potri.002G257900.1, LOC_Os10g32980.1, GSMUA_Achr5G15720_001, Glyma.09G051100.1, GSVIVT01028402001, Glyma.13G126000.1, Eucgr.A01324.1, Pavir.Ib00804.1, GRMZM2G445905_T03, PDK_30s681141g006, Migut.D00465.1, Cucsa.348940.1, hv_contig_51765_CesA4, 29637.m000754, Glyma.15G157100.1, Si034020m, Gorai.001G238100.1, Glyma.08G088400.1, Gorai.004G057400.1, Sobic.001G224300.1, evm.model.supercontig_165.53, orange1.1g001574m, AT5G44030.1, Aquca_017_00791.1, Solyc09g072820.2.1,
CesA3	Aquca_039_00062.1, Bradi2g49912.1, AT4G18780.1, Potri.011G069600.1, Glyma.05G160000.1, PDK_30s6550949g001, Migut.M01014.1, LOC_Os01g54620.1, GRMZM2G037413_T01, Sobic.003G296400.1, Gorai.011G037900.1, orange1.1g002020m, GSVIVT01021248001, Pavir.J30974.1, Eucgr.D00476.1, Glyma.08G117500.1, Gorai.009G161200.1, Si000179m, 30026.m001452, Glyma.04G063800.1, Solyc02g072240.2.1, ppa024725m, Glyma.06G065000.1, GRMZM2G175848_T01, Potri.004G059600.1, GSMUA_Achr6G31810_001, Pavir.Eb03139.1, Cucsa.212920.1, GSMUA_Achr9G05580_001, hv_contig_272477B_CesA5
CesA4	hv_contig_48990_CesA9 ,Aquca_001_00824.1 ,Migut.E01791.1 ,Pavir.J04619.1 ,Glyma.04G067900.1 ,Eucgr.J01639.1 ,Potri.006G251900.1 ,GSVIVT01035830001 ,Bradi2g34240.1 ,Eucgr.L02402.1 ,AT2G25540.1

	,Eucgr.H00939.1 ,AT4G32410.1 ,Si005742m ,ppa000611m ,Eucgr.C02801.1 ,Si021050m ,Sobic.003G049600.1 ,GSMUA_Achr2G19700_001 ,Pavir.Eb00438.1 ,Sobic.009G063400.1 ,Solyc08g061100.2.1 ,29848.m004649 ,Glyma.06G069600.1 ,GRMZM2G112336_T01 ,Pavir.Ca01073.1 ,Pavir.Ea00385.1 ,LOC_Os05g08370.1 ,Potri.018G029400.1 ,orange1.1g001399m ,Eucgr.C01769.1 ,Cucsa.310430.1 ,Si000133m ,GSMUA_Achr5G06050_001 ,Gorai.009G010200.1 ,GRMZM2G027723_T01 ,hv_contig_39637_CesA6 ,GRMZM2G104092_T01 ,evm.model.supercontig_115.8 ,GRMZM2G039454_T01 ,Gorai.009G009500.1
CesA4_A	hv_contig_48990_CesA9 ,Pavir.J04619.1 ,Bradi2g34240.1 ,Si005742m ,Si021050m ,Sobic.009G063400.1 ,GRMZM2G112336_T01 ,Pavir.Ca01073.1 ,LOC_Os05g08370.1 ,GRMZM2G027723_T01 ,hv_contig_39637_CesA6 ,GRMZM2G104092_T01
CesA4_B	Sobic.003G049600.1 ,Pavir.Eb00438.1 ,Pavir.Ea00385.1 ,Si000133m ,GRMZM2G039454_T01
CesA5	Bradi1g54250.1 ,29761.m000416 ,Solyc01g087210.2.1 ,Potri.006G052600.1 ,GSVIVT01032096001 ,Cucsa.127880.1 ,GSMUA_AchrUn_randomG09460_001 ,Potri.001G266400.1 ,Pavir.Ba03256.1 ,GRMZM2G424832_T01 ,ppa000593m ,Glyma.15G275000.1 ,GRMZM2G111642_T01 ,Sobic.001G045700.1 ,29606.m000101 ,LOC_Os07g10770.1 ,Pavir.J12858.1 ,evm.TU.contig_29394.1 ,Glyma.12G237000.1 ,GSMUA_Achr1G22920_001 ,LOC_Os03g59340.1 ,Cucsa.164240.1 ,Sobic.002G075500.1 ,Potri.016G054900.1 ,Pavir.J33961.1 ,orange1.1g001413m ,Eucgr.G03380.1 ,Aquca_014_00774.1 ,hv_contig_46016_CesA3 ,hv_contig_39703_CesA1 ,Si028770m ,GRMZM2G150404_T01 ,Gorai.004G172400.1 ,Bradi1g04597.1 ,Potri.009G060800.1 ,Glyma.13G202500.1 ,GRMZM2G018241_T01 ,Gorai.003G092600.1 ,Eucgr.J01278.1 ,Gorai.004G065900.1 ,Migut.N01600.1 ,Pavir.J35010.1 ,Si034016m ,Pavir.J11713.1 ,PDK_30s667991g005 ,Glyma.09G103000.1 ,AT5G05170.1 ,GSVIVT01033297001
CesA5_A	29761.m000416 ,Solyc01g087210.2.1 ,Potri.006G052600.1 ,GSVIVT01032096001 ,Cucsa.127880.1 ,Potri.001G266400.1 ,ppa000593m ,Glyma.15G275000.1 ,29606.m000101 ,evm.TU.contig_29394.1 ,Glyma.12G237000.1 ,Cucsa.164240.1 ,Potri.016G054900.1 ,orange1.1g001413m ,Eucgr.G03380.1 ,Aquca_014_00774.1 ,Gorai.004G172400.1 ,Potri.009G060800.1 ,Glyma.13G202500.1

	,Gorai.003G092600.1 ,Eucgr.J01278.1 ,Gorai.004G065900.1 ,Migut.N01600.1 ,Glyma.09G103000.1 ,AT5G05170.1 ,GSVIVT01033297001
CesA5_B	GSMUA_AchrUn_randomG09460_001 ,GSMUA_Achr1G22920_001 ,PDK_30s667991g005
CesA5_C	Bradi1g54250.1 ,Pavir.Ba03256.1 ,GRMZM2G424832_T01 ,LOC_Os07g10770.1 ,Sobic.002G075500.1 ,hv_contig_39703_CesA1 ,Si028770m ,GRMZM2G150404_T01 ,GRMZM2G018241_T01 ,Pavir.J35010.1/1-
CesA5_D	GRMZM2G111642_T01 ,Sobic.001G045700.1 ,Pavir.J12858.1 ,LOC_Os03g59340.1 ,Pavir.J33961.1 ,hv_contig_46016_CesA3 ,Bradi1g04597.1 ,Si034016m ,Pavir.J11713.1
CesA6	Pavir.Ba01088.1 ,Si028762m ,Eucgr.B03971.1 ,GRMZM2G113137_T01 ,LOC_Os07g14850.1 ,GRMZM2G028353_T01 ,ppa000557m ,GRMZM2G082580_T01 ,GSMUA_Achr11G10200_001 ,Sobic.002G118700.1 ,Si028766m ,Sobic.001G021500.1 ,Pavir.J01772.1 ,GSMUA_Achr9G12260_001 ,GRMZM2G177631_T01 ,orange1.1g045222m ,Pavir.J11851.1 ,Eucgr.B01532.1 ,evm.model.supercontig_36.82 ,Bradi1g02510.1 ,Pavir.J27681.1 ,LOC_Os03g62090.1 ,GSMUA_Achr7G02100_001 ,Potri.013G019800.1 ,Potri.005G027600.1 ,GSVIVT01013471001 ,GSVIVT01023837001 ,Si028764m ,hv_contig_46953_CesA2 ,Sobic.002G094600.1 ,Eucgr.B01562.1 ,Si034009m ,GSMUA_AchrUn_randomG04950_001 ,Aquca_017_00003.2 ,Bradi1g53207.1 ,Pavir.J34300.1 ,GRMZM2G025231_T02 ,Pavir.J26736.1 ,Eucgr.F03635.1 ,Bradi1g29060.3 ,LOC_Os07g24190.1
CesA6_A	Si028762m ,LOC_Os07g14850.1 ,GRMZM2G082580_T01 ,GRMZM2G177631_T01 ,Pavir.J11851.1 ,Sobic.002G094600.1 ,Bradi1g53207.1 ,Pavir.J34300.1
CesA6_B	GRMZM2G113137_T01 ,Sobic.001G021500.1 ,Pavir.J01772.1 ,Si034009m ,Pavir.J26736.1
CesA6_C	Pavir.Ba01088.1 ,GRMZM2G028353_T01 ,Sobic.002G118700.1 ,Si028766m ,Pavir.J27681.1 ,Si028764m ,GRMZM2G025231_T02
CesA7	Sobic.010G196300.1 ,orange1.1g001369m ,GSMUA_Achr8G07960_001 ,Potri.007G076500.1 ,Gorai.009G255100.1 ,Potri.005G194200.1 ,Migut.H02466.1 ,28093.m000117 ,Glyma.10G223500.1 ,Pavir.J38382.1

	,Potri.002G066600.1 ,Cucsa.325790.1 ,Aquca_023_00159.1 ,GRMZM5G840940_T01 ,ppa000559m ,orange1.1g002480m ,orange1.1g001373m ,Solyc12g056580.1.1 ,AT5G64740.1 ,AT4G39350.1 ,PDK_30s785711g002 ,Gorai.008G100200.1 ,Solyc04g071650.2.1 ,Glyma.05G187300.1 ,Pavir.Db00737.1 ,evm.model.supercontig_49.102 ,Glyma.02G080900.1 ,Eucgr.H00646.1 ,Eucgr.I00286.1 ,Glyma.16G165900.1 ,Potri.005G087500.1 ,ppa000567m ,AT2G21770.1 ,Si005779m ,AT5G09870.1 ,Glyma.08G145600.1 ,Gorai.010G134500.1 ,29863.m001055 ,Eucgr.F04212.1 ,Gorai.003G049300.1 ,Migut.C00978.1 ,GSMUA_Achr8G32810_001 ,evm.model.supercontig_72.37 ,GSVIVT01034552001 ,GSVIVT01022285001 ,Migut.N03206.1 ,Eucgr.H02200.1 ,Gorai.002G150300.1 ,Solyc11g005560.1.1 ,Cucsa.229470.1
CesA7_A	orange1.1g001369m ,Potri.007G076500.1 ,Migut.H02466.1 ,28093.m000117 ,Aquca_023_00159.1 ,AT5G64740.1 ,AT4G39350.1 ,Gorai.008G100200.1 ,Glyma.05G187300.1 ,Eucgr.I00286.1 ,Potri.005G087500.1 ,ppa000567m ,AT2G21770.1 ,AT5G09870.1 ,Glyma.08G145600.1 ,Gorai.003G049300.1 ,evm.model.supercontig_72.37 ,GSVIVT01022285001 ,Gorai.002G150300.1 ,Solyc11g005560.1.1 ,Cucsa.229470.1
CesA7_B	Gorai.009G255100.1 ,Potri.005G194200.1 ,Glyma.10G223500.1 ,Potri.002G066600.1 ,Cucsa.325790.1 ,ppa000559m ,orange1.1g002480m ,orange1.1g001373m ,Solyc12g056580.1.1 ,Solyc04g071650.2.1 ,evm.model.supercontig_49.102 ,Glyma.02G080900.1 ,Eucgr.H00646.1 ,Glyma.16G165900.1 ,Gorai.010G134500.1 ,29863.m001055 ,Eucgr.F04212.1 ,Migut.C00978.1 ,GSVIVT01034552001 ,Migut.N03206.1 ,Eucgr.H02200.1
CslD1	Sobic.001G283400.1 ,Glyma.01G232500.1 ,LOC_Os06g02180.1 ,orange1.1g003243m ,GSMUA_AchrUn_randomG07850_001 ,Eucgr.D02228.1 ,30073.m002256 ,Glyma.09G208200.1 ,Si005721m ,Gorai.004G257300.1 ,Gorai.003G052200.1 ,Migut.E00738.1 ,Bradi3g34490.2 ,Solyc08g005280.1.1 ,AT1G32180.1 ,orange1.1g001213m ,hv_contig_41777_CSL_D2 ,Si033974m ,GRMZM5G870176_T01 ,Glyma.11G010400.1 ,Pavir.Db02441.1 ,ppa021772m ,Gorai.008G223700.1 ,Cucsa.017540.1 ,AT3G03050.1 ,Eucgr.E00226.1 ,Potri.013G082200.1 ,Glyma.01G014000.1 ,PDK_30s693711g002 ,Aquca_005_00576.1 ,LOC_Os10g42750.1 ,Pavir.Ib03058.1 ,Solyc08g076320.2.1 ,30068.m002658 ,Bradi1g50170.1 ,Migut.F00094.1 ,Migut.F00941.1 ,GSMUA_Achr1G23960_001 ,GRMZM2G436299_T01 ,ppa000493m ,Solyc03g097050.2.1 ,Sobic.010G008600.1 ,Cucsa.017530.1 ,Pavir.J22394.1 ,AT5G16910.1 ,ppa000473m ,GSMUA_Achr11G23440_001

	,Potri.003G097100.1 ,Potri.001G136200.1 ,hv_contig_2547805_CSL_D3 ,Pavir.Ia02691.1 ,Eucgr.K00085.1 ,Potri.019G049700.1 ,Glyma.12G017600.1
CslD1_A	orange1.1g003243m ,Eucgr.D02228.1 ,Glyma.09G208200.1 ,AT1G32180.1 ,ppa021772m ,Gorai.008G223700.1 ,Eucgr.E00226.1 ,Glyma.01G014000.1 ,30068.m002658 ,ppa000493m ,Potri.003G097100.1 ,Potri.001G136200.1
CslD1_B	Glyma.01G232500.1 ,30073.m002256 ,Gorai.004G257300.1 ,Gorai.003G052200.1 ,Migut.E00738.1 ,Solyc08g005280.1.1 ,orange1.1g001213m ,Glyma.11G010400.1 ,Cucsa.017540.1 ,AT3G03050.1 ,Potri.013G082200.1 ,Aquca_005_00576.1 ,Solyc08g076320.2.1 ,Migut.F00094.1 ,Migut.F00941.1 ,Solyc03g097050.2.1 ,Cucsa.017530.1 ,AT5G16910.1 ,ppa000473m ,Eucgr.K00085.1 ,Potri.019G049700.1 ,Glyma.12G017600.1
CslD1_C	LOC_Os06g02180.1 ,Si005721m ,hv_contig_41777_CSL_D2 ,GRMZM5G870176_T01 ,Pavir.Db02441.1 ,Bradi1g50170.1 ,Sobic.010G008600.1 ,Pavir.J22394.1
CslD1_D	Sobic.001G283400.1 ,Bradi3g34490.2 ,Si033974m ,PDK_30s693711g002 ,LOC_Os10g42750.1 ,Pavir.Ib03058.1 ,GRMZM2G436299_T01 ,hv_contig_2547805_CSL_D3 ,Pavir.Ia02691.1
CslD2	Sobic.008G125700.1 ,Cucsa.096870.1 ,evm.model.supercontig_151.45 ,Cucsa.096880.1 ,GSVIVT01028071001 ,orange1.1g001071m ,Pavir.Ca02654.1 ,Migut.D00114.1 ,Cucsa.096890.1 ,29986.m001674 ,AT1G02730.1 ,Aquca_003_00613.1 ,Si021011m ,PDK_30s1109201g002 ,Pavir.Cb00380.1 ,hv_contig_1576831_CSL_D4 ,Bradi4g05027.1 ,GRMZM2G015886_T01 ,Solyc09g075550.2.1 ,GSMUA_Achr3G24160_001 ,Potri.014G125100.1 ,Glyma.03G217500.1 ,Gorai.008G142900.1 ,Potri.002G200300.1 ,LOC_Os12g36890.1
CslD3	Aquca_026_00396.1 ,Eucgr.H00079.1 ,Pavir.Db01215.1 ,GSVIVT01021798001 ,Sobic.010G146000.1 ,ppa000644m ,29904.m002970 ,Pavir.Da01737.1 ,Glyma.09G119000.1 ,LOC_Os06g22980.1 ,Solyc05g053560.2.1 ,Potri.003G177800.1 ,Bradi2g03380.1 ,Potri.001G050200.1 ,Migut.G00645.1 ,AT2G33100.1 ,orange1.1g036064m ,hv_contig_6299_CSL_D6 ,Si008500m ,GRMZM2G061764_T01 ,Gorai.006G220600.1
CslD4	Pavir.Fa01390.1 ,Pavir.J37667.1 ,Potri.009G170000.1 ,ppa000490m ,GRMZM2G044269_T01 ,Eucgr.H05010.1 ,evm.model.supercontig_426.3 ,Gorai.012G137800.1 ,PDK_30s1009171g003 ,Migut.F02137.1

	,30162.m001289 ,AT4G38190.1 ,Si015820m ,Bradi3g22345.1 ,Sobic.007G100800.1 ,Cucsa.099960.1 ,Glyma.02G286100.1 ,orange1.1g042084m ,Solyc01g067520.2.1 ,Solyc10g074620.1.1 ,Glyma.14G029200.1 ,GSVIVT01023850001 ,GSMUA_Achr2G05580_001 ,hv_contig_136793_CSL_D1 ,LOC_Os08g25710.1 ,Potri.004G208800.1 ,Aquca_007_00954.1
CslF1	Si013204m ,Bradi3g16307.1 ,GRMZM2G110145_T01 ,LOC_Os08g06380.1 ,hv_contig_41513_CSL_F6 ,Pavir.Fb00422.1 ,GRMZM2G122277_T01 ,Sobic.007G050600.1
CslF2	Sobic.001G242000.1 ,LOC_Os10g20260.1 ,Si034259m ,Pavir.Ba01324.1 ,GRMZM2G164761_T01 ,hv_contig_36872_CSL_F7
CslF3	Bradi1g25117.1 ,hv_contig_6524_CSL_F4 ,LOC_Os07g36740.1 ,Pavir.Bb02875.1 ,LOC_Os07g36700.1 ,Pavir.J22621.1 ,LOC_Os07g36690.1 ,Si028885m ,Sobic.002G333900.1 ,Pavir.Ba00686.1 ,GRMZM2G103972_T01 ,hv_contig_1585560_F11
CslF4	Pavir.Bb03029.1 ,Sobic.002G333800.1 ,Bradi1g25107.1 ,Pavir.Bb03028.1 ,hv_contig_37718_CSL_F8 ,GRMZM2G113432_T01 ,Pavir.Ba00687.1 ,LOC_Os07g36630.1 ,Si028860m
CslF5	Bradi3g45515.1 ,LOC_Os07g36610.1 ,Pavir.Bb03027.1 ,Pavir.Ba00688.1 ,hv_MLOC_7825_CSL_F12 ,Si028873m ,Sobic.002G334300.1 ,hv_contig_43489_CSL_F9
CslF6	hv_contig_1565725_CSL_F10 ,Sobic.002G171200.1 ,Bradi1g25157.2 ,Bradi1g25150.1 ,Sobic.002G334500.1 ,Pavir.Ba00685.2 ,Sobic.002G334400.1 ,Pavir.J07223.1 ,Pavir.Ba00685.1
CslF7	Si031960m ,Si032230m ,Pavir.Bb03031.1 ,Sobic.002G334000.1 ,GRMZM2G339645_T01 ,Sobic.002G334200.1 ,hv_contig_43435_CSL_F3 ,Sobic.002G334100.1 ,Bradi1g25130.1 ,LOC_Os07g36750.1 ,GRMZM2G367267_T01
CslB	Gorai.007G171800.1 ,Eucgr.K00779.1 ,Glyma.12G192100.1 ,Glyma.12G096900.1 ,GSVIVT01030468001 ,orange1.1g006357m ,Cucsa.148680.1 ,Glyma.12G191800.1 ,GSVIVT01030462001 ,Potri.002G227300.1 ,Cucsa.309270.1 ,Aquca_011_00013.1 ,GSVIVT01030458001 ,Cucsa.072550.1 ,GSVIVT01030461001 ,Cucsa.309280.1 ,Eucgr.K00778.1 ,Glyma.12G192000.1 ,Aquca_070_00023.1

	,Eucgr.K00782.1 ,AT2G32530.1 ,AT4G15290.1 ,Solyc07g051820.2.1 ,ppa026099m ,AT2G32540.1 ,GSVIVT01030464001 ,Potri.014G155300.1 ,29901.m000406 ,Cucsa.072580.1 ,AT2G32610.1 ,orange1.1g007647m ,Glyma.12G191700.1 ,Glyma.06G307900.1 ,orange1.1g006639m ,29901.m000405 ,ppa001889m ,Cucsa.309300.1 ,Cucsa.148670.1 ,Glyma.13G310300.1 ,AT4G15320.1 ,Eucgr.K00781.1 ,GSVIVT01030467001 ,AT2G32620.1 ,ppa001909m ,GSVIVT01030456001 ,Glyma.12G191500.1
CslH	Sobic.006G080700.1 ,LOC_Os04g35020.1 ,Sobic.006G080800.1 ,Si009413m ,Pavir.Gb01691.1 ,hv_contig_37984_CSL_H1 ,LOC_Os04g35030.1 ,LOC_Os10g20090.1 ,Pavir.Ga01895.1 ,Si016419m ,Sobic.006G080600.1 ,Bradi5g10130.1 ,Pavir.Da01545.1 ,GSMUA_Achr9G08770_001 ,GRMZM2G074546_T02
CslE1	Gorai.002G101600.1 ,Migut.M01644.1 ,Aquca_118_00010.1 ,orange1.1g009753m ,evm.model.supercontig_12.121 ,29629.m001411 ,Solyc12g015770.1.1 ,Migut.H01750.1 ,ppa001952m ,Cucsa.247810.1 ,Glyma.14G012800.1 ,orange1.1g006553m ,Glyma.02G301200.1 ,GSVIVT01014705001
CslE2	GRMZM2G014558_T01 ,Eucgr.E03851.1 ,orange1.1g005037m ,Si029066m ,GSVIVT01006761001 ,GSVIVT01004458001 ,orange1.1g009524m ,Solyc07g065660.2.1 ,Aquca_093_00036.1 ,GSMUA_Achr4G13080_001 ,Potri.006G004300.1 ,GSVIVT01007039001 ,Bradi4g33080.1 ,Glyma.08G330700.1 ,Gorai.013G193300.1 ,GSVIVT01014703001 ,Pavir.Ba01587.1 ,Migut.H01751.1 ,LOC_Os09g30120.1 ,Aquca_012_00081.1 ,Sobic.002G237900.1 ,GSVIVT01007042001 ,PDK_30s1035221g001 ,Si016995m ,ppa018204m ,ppa001941m ,Glyma.08G330600.1 ,GSVIVT01007038001 ,Aquca_093_00061.1 ,LOC_Os02g49332.1 ,Gorai.007G261800.1 ,LOC_Os09g30130.1 ,Aquca_093_00038.1 ,GRMZM2G012044_T01 ,Gorai.013G193400.1 ,Cucsa.210990.1 ,GSVIVT01007043001 ,evm.model.supercontig_27.221 ,Glyma.14G012900.1 ,hv_contig_275250_CSL_E2 ,29629.m001410 ,GSVIVT01006760001 ,Aquca_112_00025.1 ,Eucgr.L03518.1 ,Bradi3g56440.1 ,GSVIVT01006766001 ,Sobic.004G255200.1 ,AT1G55850.1 ,Pavir.Aa00687.1 ,Gorai.002G101400.1 ,Eucgr.H05074.1 ,GSMUA_Achr9G28000_001 ,GSVIVT01006764001 ,Potri.001G369100.1 ,29629.m001409 ,Cucsa.247820.1 ,Eucgr.F03681.1 ,Sobic.002G238300.1 ,Bradi4g33090.1 ,Gorai.007G261500.1 ,Eucgr.E03849.1 ,Si029057m ,Aquca_001_00717.1 ,GSMUA_Achr9G18830_001 ,Pavir.J40103.1 ,ppa001936m ,Gorai.007G261700.1 ,Aquca_112_00031.1 ,Eucgr.E03847.1 ,GSVIVT01006763001 ,Gorai.007G261300.1

	,Potri.006G004200.1 ,hv_contig_50865_CSL_E ,Pavir.Bb02446.1 ,Eucgr.E03846.1 ,Si029554m ,GSVIVT01007333001 ,Pavir.Ab02815.1
CslE2_A	Eucgr.E03851.1 ,GSVIVT01006761001 ,GSVIVT01004458001 ,Potri.006G004300.1 ,GSVIVT01007039001 ,Gorai.013G193300.1 ,GSVIVT01007042001 ,GSVIVT01007038001 ,Gorai.007G261800.1 ,Gorai.013G193400.1 ,GSVIVT01007043001 ,evm.model.supercontig_27 ,GSVIVT01006760001 ,Eucgr.L03518.1 ,GSVIVT01006766001 ,Eucgr.H05074.1 ,GSVIVT01006764001 ,Gorai.007G261500.1 ,Eucgr.E03849.1 ,Gorai.007G261700.1 ,Eucgr.E03847.1 ,GSVIVT01006763001 ,Gorai.007G261300.1 ,Potri.006G004200.1 ,Eucgr.E03846.1 ,GSVIVT01007333001
CslE2_B	orange1.1g005037m ,orange1.1g009524m ,Solyc07g065660.2.1 ,Aquca_093_00036.1 ,Glyma.08G330700.1 ,GSVIVT01014703001 ,Migut.H01751.1 ,Aquca_012_00081.1 ,ppa018204m ,ppa001941m ,Glyma.08G330600.1 ,Aquca_093_00061.1 ,Aquca_093_00038.1 ,Cucsa.210990.1 ,Glyma.14G012900.1 ,29629.m001410 ,Aquca_112_00025.1 ,AT1G55850.1 ,Gorai.002G101400.1 ,Potri.001G369100.1 ,29629.m001409 ,Cucsa.247820.1 ,Eucgr.F03681.1 ,Aquca_001_00717.1 ,ppa001936m ,Aquca_112_00031.1
CslE2_C	PDK_30s1035221g001 ,Si016995m ,LOC_Os02g49332.1 ,GRMZM2G012044_T01 ,Bradi3g56440.1 ,Sobic.004G255200.1 ,Pavir.Aa00687.1 ,GSMUA_Achr9G28000_001 ,Pavir.Ab02815.1
CslE2_D	GRMZM2G014558_T01 ,Si029066m ,Bradi4g33080.1 ,Pavir.Ba01587.1 ,LOC_Os09g30120.1 ,Sobic.002G237900.1 ,LOC_Os09g30130.1 ,hv_contig_275250_CSL_E2 ,Sobic.002G238300.1 ,Bradi4g33090.1 ,Si029057m ,Pavir.J40103.1 ,hv_contig_50865_CSL_E_INCOMP ,Pavir.Bb02446.1 ,Si029554m
CslG	30068.m002517 ,30068.m002518 ,30068.m002519 ,AT4G23990.1 ,AT4G24000.1 ,AT4G24010.1 ,Aquca_037_00085.1 ,Aquca_037_00086.1 ,Cucsa.114770.1 ,Cucsa.114810.1 ,Cucsa.114830.1 ,Eucgr.D01765.1 ,Eucgr.D01766.1 ,Eucgr.D01768.1 ,GSMUA_Achr8G05340_001 ,GSMUA_AchrG23890_001 ,GSVIVT01019568001 ,GSVIVT01019570001 ,GSVIVT01019572001 ,GSVIVT01019575001 ,GSVIVT01019580001 ,GSVIVT01019581001 ,GSVIVT01019582001 ,GSVIVT01019583001 ,GSVIVT01019586001 ,GSVIVT01019587001 ,GSVIVT01019589001 ,GSVIVT01019591001 ,Gorai.006G206600.1 ,Gorai.006G206700.1 ,Gorai.006G206800.1 ,Gorai.008G256300.1 ,Migut.A00610.1 ,Migut.A01179.1

	,Pavir.Aa00002.1 ,Pavir.Ea01504.1 ,Potri.003G142300.1 ,Potri.003G142400.1 ,Potri.003G142500.1 ,Solyc08g082640.2.1 ,Solyc08g082660.2.1 ,Solyc08g082670.2.1 ,Solyc12g014430.1.1 ,evm.model.supercontig_15.5 ,orange1.1g045732m ,ppa001971m ,ppa001978m ,ppb017427m
CslJ	Sobic.003G442500.1 ,hv_contig_1593432X_CSL_J ,Pavir.Eb04036.1 ,GRMZM2G122431_T01 ,Si021430m ,Si024965m
CslM1	Potri.010G074700.1 ,Eucgr.E00820.1 ,GSVIVT01018143001 ,Eucgr.H00186.1 ,orange1.1g004562m ,orange1.1g004779m ,Gorai.006G247400.1 ,Eucgr.H00188.1 ,Glyma.06G324300.1 ,ppa001935m ,Migut.G00447.1 ,Solyc03g005450.2.1 ,Glyma.13G174300.1 ,Eucgr.H00189.1 ,Potri.010G074800.1 ,GSVIVT01027716001 ,Migut.G00456.1 ,29603.m000538 ,Eucgr.E00819.1 ,Gorai.011G167800.1 ,ppa001861m ,Eucgr.E00821.1 ,29603.m000539 ,Gorai.005G047100.1 ,orange1.1g004695m ,GSVIVT01027717001 ,29603.m000541 ,Gorai.005G047700.1 ,Gorai.006G247300.1 ,Cucsa.255110.1 ,Glyma.04G255400.1 ,orange1.1g004752m ,GSVIVT01018144001 ,Aquca_017_00670.1 ,Gorai.006G247500.1 ,Glyma.11G151800.1 ,ppa001867m ,orange1.1g004692m ,Glyma.10G189300.1 ,orange1.1g004688m ,Aquca_017_00672.1 ,Solyc00g030000.1.1
CslM1_A	Glyma.06G324300.1 ,Migut.G00447.1 ,Gorai.011G167800.1 ,Glyma.04G255400.1 ,Glyma.11G151800.1 ,Solyc00g030000.1.1
CslM1_B	Potri.010G074700.1 ,Eucgr.E00820.1 ,GSVIVT01018143001 ,Eucgr.H00186.1 ,orange1.1g004562m ,orange1.1g004779m ,Gorai.006G247400.1 ,Eucgr.H00188.1 ,ppa001935m ,Solyc03g005450.2.1 ,Glyma.13G174300.1 ,Eucgr.H00189.1 ,Potri.010G074800.1 ,GSVIVT01027716001 ,Migut.G00456.1 ,29603.m000538 ,Eucgr.E00819.1 ,ppa001861m ,Eucgr.E00821.1 ,29603.m000539 ,Gorai.005G047100.1 ,orange1.1g004695m ,GSVIVT01027717001 ,29603.m000541 ,Gorai.005G047700.1 ,Gorai.006G247300.1 ,Cucsa.255110.1 ,orange1.1g004752m ,GSVIVT01018144001 ,Gorai.006G247500.1 ,ppa001867m ,orange1.1g004692m ,Glyma.10G189300.1 ,orange1.1g004688m
CslM1_C	Potri.010G074700.1 ,Solyc03g005450.2.1 ,29603.m000538 ,GSVIVT01027717001 ,Gorai.005G047700.1
CslA1	Bradi4g38970.1 ,Hv_CslA6 ,Pavir.Bb01546.1 ,Si032179m ,GRMZM2G107754_T02 ,Sobic.002G139900.1 ,GSMUA_AchrUn_randomT04290_001 ,Glyma.03G183500.1

	,Glyma.19G184200.1 ,Aquca_030_00077.1 ,Migut.L01705.1 ,Migut.J00655.1 ,orange1.1g044519m ,GSVIVT01025737001
CslA2	Migut.N02519.1 ,Migut.D02330.1 ,AT3G56000_CSL_A14 ,AT2G35650_CSL_A7 ,AT4G16590_CSL_A1 ,AT1G24070_CSL_A10 ,AT4G13410_CSL_A15 ,AT5G16190_CSL_A11 ,AT1G23480_CSL_A3 ,Si016912m ,Sobic.004G075900.1 ,GRMZM2G105631_T01 ,LOC_Os02g09930.1 ,Bradi3g06740.1 ,hv_CSL_A2 ,Cucsa.378620.1 ,Migut.N02518.1 ,Migut.D02331.1 ,Solyc11g066820.1.1 ,Solyc06g074630.2.1 ,Glyma.13G319600.1 ,Glyma.11G189600.1 ,Glyma.12G084600.1 ,GSVIVT01031405001 ,AT5G22740_CSL_A2 ,Eucgr.J00420.1 ,Gorai.004G008100.1 ,29991.m000638 ,Potri.004G189000.1 ,Potri.009G149700.1 ,Cucsa.385420.1 ,ppa004330m ,ppa003918m ,evm.model.supercontig_47.50 ,Gorai.006G109100.1 ,Solyc11g007600.1.1 ,Aquca_019_00048.1 ,Aquca_019_00047.1 ,Bradi1g35647.1 ,Si008033m ,Pavir.Db00724.1 ,Sobic.010G197300.1 ,GRMZM2G334142_CSL_incomp ,GRMZM2G443715_T01 ,LOC_Os06g42020.1 ,orange1.1g041333m ,GSMUA_Achr6T07540_001 ,Solyc05g055410.1.1 ,Cucsa.083070.1 ,Cucsa.083080.1 ,GSMUA_Achr5T15920_001 ,GSMUA_Achr10T11180_001 ,GSMUA_Achr2T20260_001 ,GSMUA_Achr6T13400_001 ,GSMUA_Achr10T15450_001 ,Glyma.13G150700.1 ,Glyma.10G065600.1 ,ppa004037m ,AT5G03760_CSL_A9 ,Migut.J00497.1 ,Glyma.19G190600.1 ,Glyma.03G190200.1 ,Solyc10g083670.1.1 ,Gorai.002G243400.1 ,Eucgr.A01558.1 ,Potri.006G116900.1 ,29092.m000444 ,Cucsa.163310.1 ,GSVIVT01033767001 ,Aquca_009_00991.1 ,Eucgr.G02715.1 ,Glyma.10G201700.1 ,Glyma.20G188600.1 ,Gorai.009G305600.1 ,Gorai.011G232100.1 ,GSVIVT01034719001 ,orange1.1g009761m ,ppa024741m ,ppa004315m ,29428.m000330 ,Potri.008G026400.1 ,Potri.010G234100.1
CslA2_A	Migut.N02519.1 ,Migut.D02330.1 ,AT3G56000_CSL_A14 ,AT2G35650_CSL_A7 ,AT4G16590_CSL_A1 ,AT1G24070_CSL_A10 ,AT4G13410_CSL_A15 ,AT5G16190_CSL_A11 ,AT1G23480_CSL_A3 ,Si016912m ,Sobic.004G075900.1 ,GRMZM2G105631_T01 ,LOC_Os02g09930.1 ,Bradi3g06740.1 ,hv_CSL_A2 ,Cucsa.378620.1 ,Migut.N02518.1 ,Migut.D02331.1 ,Solyc11g066820.1.1 ,Solyc06g074630.2.1 ,Glyma.13G319600.1 ,Glyma.11G189600.1 ,Glyma.12G084600.1 ,GSVIVT01031405001 ,AT5G22740_CSL_A2 ,Eucgr.J00420.1 ,Gorai.004G008100.1 ,29991.m000638 ,Potri.004G189000.1 ,Potri.009G149700.1 ,Cucsa.385420.1 ,ppa004330m ,ppa003918m ,evm.model.supercontig_47.50 ,Gorai.006G109100.1 ,Solyc11g007600.1.1 ,Aquca_019_00048.1 ,Aquca_019_00047.1 ,Bradi1g35647.1 ,Si008033m

	,Pavir.Db00724.1 ,Sobic.010G197300.1 ,GRMZM2G334142_CSL_incomp ,GRMZM2G443715_T01 ,LOC_Os06g42020.1 ,orange1.1g041333m ,GSMUA_Achr6T07540_001 ,Solyc05g055410.1.1 ,Cucsa.083070.1 ,Cucsa.083080.1 ,GSMUA_Achr5T15920_001 ,GSMUA_Achr10T11180_001 ,GSMUA_Achr2T20260_001 ,GSMUA_Achr6T13400_001 ,GSMUA_Achr10T15450_001 ,Glyma.13G150700.1 ,Glyma.10G065600.1 ,ppa004037m ,AT5G03760_CSL_A9 ,Migut.J00497.1 ,Glyma.19G190600.1 ,Glyma.03G190200.1 ,Solyc10g083670.1.1 ,Gorai.002G243400.1 ,Eucgr.A01558.1 ,Potri.006G116900.1 ,29092.m000444 ,Cucsa.163310.1 ,GSVIVT01033767001 ,Aquca_009_00991.1 ,Eucgr.G02715.1 ,Glyma.10G201700.1 ,Glyma.20G188600.1 ,Gorai.009G305600.1 ,Gorai.011G232100.1 ,GSVIVT01034719001 ,orange1.1g009761m ,ppa024741m ,ppa004315m ,29428.m000330 ,Potri.008G026400.1 ,Potri.010G234100.1
CslA2_B	Si016912m ,Sobic.004G075900.1 ,GRMZM2G105631_T01 ,LOC_Os02g09930.1 ,Bradi3g06740.1 ,hv_CSL_A2 ,Bradi1g35647.1 ,Si008033m ,Pavir.Db00724.1 ,Sobic.010G197300.1 ,GRMZM2G334142_CSL_incomp ,GRMZM2G443715_T01 ,LOC_Os06g42020.1 ,GSMUA_Achr6T07540_001 ,GSMUA_Achr5T15920_001 ,GSMUA_Achr10T11180_001 ,GSMUA_Achr2T20260_001 ,GSMUA_Achr6T13400_001 ,GSMUA_Achr10T15450_001
CslA2_C	Migut.N02519.1 ,Migut.D02330.1 ,Cucsa.378620.1 ,Migut.N02518.1 ,Migut.D02331.1 ,Solyc11g066820.1.1 ,Solyc06g074630.2.1 ,Glyma.13G319600.1 ,Glyma.11G189600.1 ,Glyma.12G084600.1 ,GSVIVT01031405001 ,AT5G22740_CSL_A2 ,Eucgr.J00420.1 ,Gorai.004G008100.1 ,29991.m000638 ,Potri.004G189000.1 ,Potri.009G149700.1 ,Cucsa.385420.1 ,ppa004330m ,ppa003918m ,evm.model.supercontig_47.50 ,Gorai.006G109100.1
CslA2_D	hv_CSL_A3 ,Bradi3g25658.1 ,LOC_Os10g26630.1 ,GRMZM2G115772_T01 ,Sobic.001G252700.1 ,Si035204m ,Pavir.Ib03508.1 ,Pavir.Ia02025.1
CslA3	Bradi3g59447.1 ,LOC_Os02g51060.1 ,Si016839m ,GRMZM2G405567_T02 ,Sobic.004G238700.1 ,LOC_Os03g26044.1 ,Pavir.Ia03178.1 ,Pavir.Ib01792.1 ,Si034919m ,LOC_Os07g43710.1 ,Si029676m ,Pavir.Bb03416.1 ,Pavir.Ba00356.1 ,GRMZM2G010142_T02 ,Sobic.002G385800.1 ,hv_CSL_A1_w ,Bradi1g20500.1 ,GSMUA_Achr6T02600_001 ,GSMUA_Achr11T09120_001 ,GSMUA_Achr3T01830_001 ,GSMUA_Achr7T21070_001 ,GSMUA_Achr7T18150_001 ,Bradi3g36697.1

	,hv_CSL_A4_w ,LOC_Os08g33740.1 ,GRMZM2G099088_T01 ,GRMZM2G108600_T03 ,Sobic.007G137400.1 ,Si015800m ,Pavir.Fa00887.1 ,Pavir.J06413.1 ,LOC_Os09g26770.2 ,LOC_Os06g12460.1 ,Bradi1g45125.1 ,Pavir.Da01284.1 ,Si006161m ,Sobic.010G093500.1 ,GRMZM2G020742_T03 ,GSMUA_Achr10T18780_001 ,hv_CSL_A3 ,Bradi3g25658.1 ,LOC_Os10g26630.1 ,GRMZM2G115772_T01 ,Sobic.001G252700.1 ,Si035204m ,Pavir.Ib03508.1 ,Pavir.Ia02025.1 ,LOC_Os03g07350.1 ,Pavir.Ib00445.1 ,Pavir.Ia04316.1 ,Sobic.001G490000.1 ,GRMZM2G178880_T01
CslA3_A	Bradi3g59447.1 ,LOC_Os02g51060.1 ,Si016839m ,GRMZM2G405567_T02 ,Sobic.004G238700.1 ,GSMUA_Achr6T02600_001 ,GSMUA_Achr11T09120_001 ,GSMUA_Achr3T01830_001 ,GSMUA_Achr7T21070_001 ,GSMUA_Achr7T18150_001
CslA3_B	Bradi3g36697.1 ,hv_CSL_A4_w ,LOC_Os08g33740.1 ,GRMZM2G099088_T01 ,GRMZM2G108600_T03 ,Sobic.007G137400.1 ,Si015800m ,Pavir.Fa00887.1 ,Pavir.J06413.1
CslA3_C	LOC_Os06g12460.1 ,Bradi1g45125.1 ,Pavir.Da01284.1 ,Si006161m ,Sobic.010G093500.1 ,GRMZM2G020742_T03
CslC1	orange1.1g041566m ,Migut.D02479.1 ,Solyc09g057640.2.1 ,AT3G07330_CSL_C6 ,Aquca_058_00217.1 ,Gorai.N005100.1 ,Cucsa.387710.1 ,Potri.002G248400.1 ,ppa002311m ,evm.model.supercontig_8.105 ,Glyma.03G086600.1 ,Glyma.16G087600.1 ,GSMUA_Achr6T27800_001
CslC2	Solyc12g088240.1.1 ,Gorai.010G028300.1 ,AT4G07960_CSL_C12 ,hv_CSL_C1 ,hv_CSL_C4 ,Bradi2g20141.1 ,GRMZM2G074792_T01 ,LOC_Os05g43530.1 ,Sobic.009G194200.1 ,Pavir.J11046.1 ,Si021376m ,Pavir.J17316.1 ,GRMZM2G454081_CSL_C_q ,Bradi2g50967.1 ,LOC_Os01g56130.1 ,GRMZM2G027794_T01 ,Sobic.003G308100.1 ,Si000541m ,Pavir.Eb03273.1 ,Pavir.Ea02824.1 ,Eucgr.F00101.1 ,GSMUA_Achr6T30200_001 ,GSMUA_Achr4T23120_001 ,GSMUA_Achr7T00750_001 ,GSMUA_Achr10T08330_001 ,GSMUA_Achr4T32270_001 ,GSMUA_Achr5T01920_001 ,Sobic.002G022700.1 ,orange1.1g005700m ,Bradi1g57552.1 ,Bradi1g07277.1 ,AC183932.3_FGT007 ,hv_CSL_C2_w ,LOC_Os03g56060.1 ,GRMZM2G028286_T01 ,Sobic.001G075600.1 ,Si034503m ,Pavir.Ia00426.1 ,Pavir.J18214.1 ,LOC_Os07g03260.1 ,Si032804m ,Pavir.Ba03910.1 ,Pavir.Bb00218.1 ,Solyc04g077470.2.1 ,Cucsa.395390.1 ,Migut.N02800.1

	,Aquca_017_00107.1 ,Migut.C00443.1 ,Glyma.14G090000.1 ,Glyma.04G048100.1 ,Glyma.06G049400.1 ,Gorai.005G098600.1 ,Gorai.009G277600.1 ,Gorai.009G222300.1 ,Aquca_083_00103.2 ,Eucgr.F02219.1 ,GSVIVT01009290001 ,30170.m013660 ,Potri.002G114200.1 ,Potri.005G146900.1 ,ppa002209m ,Cucsa.353070.1
CslC2_A	Sobic.002G022700.1 ,Bradi1g57552.1 ,Bradi1g07277.1 ,AC183932.3_FGT007 ,hv_CSL_C2_w ,LOC_Os03g56060.1 ,GRMZM2G028286_T01 ,Sobic.001G075600.1 ,Si034503m ,Pavir.Ia00426.1 ,Pavir.J18214.1 ,LOC_Os07g03260.1 ,Si032804m ,Pavir.Ba03910.1 ,Pavir.Bb00218.1
CslC2_B	hv_CSL_C1 ,Bradi2g20141.1 ,GRMZM2G074792_T01 ,LOC_Os05g43530.1 ,Sobic.009G194200.1 ,Pavir.J11046.1 ,Si021376m ,Pavir.J17316.1
CslC2_C	hv_CSL_C4 ,GRMZM2G454081_CSL_C_q ,Bradi2g50967.1 ,LOC_Os01g56130.1 ,GRMZM2G027794_T01 ,Sobic.003G308100.1 ,Si000541m ,Pavir.Eb03273.1 ,Pavir.Ea02824.1
CslC3	Aquca_002_00287.1 ,Solyc02g089640.2.1 ,Migut.H00151.1 ,Cucsa.158900.1 ,AT3G28180_CSL_C4 ,Glyma.13G070300.1 ,Glyma.19G012700.1 ,ppa002511m ,Gorai.006G028400.1 ,29822.m003385 ,orange1.1g006104m ,Si029148m ,GRMZM2G173759_T01 ,Sobic.002G208200.1 ,Eucgr.I01833.1 ,LOC_Os09g25900.1 ,Bradi3g19087.1 ,hv_CSL_C3 ,GSMUA_Achr3T28000_001 ,GSMUA_Achr4T05880_001 ,Solyc08g006310.2.1 ,LOC_Os08g15420.1 ,GSVIVT01032523001 ,Pavir.Fb01089.1 ,Pavir.Fa01312.1 ,Si013296m ,Sobic.007G090600.1 ,GRMZM2G135286_T01 ,GRMZM2G142685_T01 ,Migut.A00427.1 ,AT2G24630.1 ,GSMUA_Achr7T19500_001 ,AT4G31590_CSL_C5 ,Glyma.14G136900.1 ,Glyma.17G196700.1 ,Cucsa.311130.1 ,Gorai.010G120900.1 ,GSVIVT01033168001 ,GSMUA_Achr6T01880_001 ,ppa002275m ,orange1.1g005507m ,Potri.006G270900.1 ,Potri.018G009300.1 ,Gorai.001G040500.1 ,Gorai.009G066500.1 ,Glyma.07G003800.1 ,Glyma.08G222800.1 ,Glyma.04G076500.1 ,Glyma.06G077700.1 ,Eucgr.C02007.1 ,29848.m004477 ,ppa002254m ,evm.model.supercontig_23.38 ,Aquca_004_00030.1 ,GSVIVT01016135001
CslC3_A	Si029148m ,GRMZM2G173759_T01 ,Sobic.002G208200.1 ,LOC_Os09g25900.1

CslC3_B	Bradi3g19087.1 ,hv_CSL_C3 ,LOC_Os08g15420.1 ,Pavir.Fb01089.1 ,Pavir.Fa01312.1 ,Si013296m ,Sobic.007G090600.1 ,GRMZM2G135286_T01 ,GRMZM2G142685_T01
CslC3_C	Solyc08g006310.2.1 ,Migut.A00427.1 ,AT2G24630.1 ,AT4G31590_CSL_C5 ,Glyma.14G136900.1 ,Glyma.17G196700.1 ,Cucsa.311130.1 ,Gorai.010G120900.1 ,GSVIVT01033168001 ,ppa002275m ,orange1.1g005507m ,Potri.006G270900.1 ,Potri.018G009300.1 ,Gorai.001G040500.1 ,Gorai.009G066500.1 ,Glyma.07G003800.1 ,Glyma.08G222800.1 ,Glyma.04G076500.1 ,Glyma.06G077700.1 ,Eucgr.C02007.1 ,29848.m004477 ,ppa002254m ,evm.model.supercontig_23.38 ,Aquca_004_00030.1 ,GSVIVT01016135001
CslC3_D	Aquca_002_00287.1 ,Solyc02g089640.2.1 ,Migut.H00151.1 ,Cucsa.158900.1 ,AT3G28180_CSL_C4 ,Glyma.13G070300.1 ,Glyma.19G012700.1 ,ppa002511m ,Gorai.006G028400.1 ,29822.m003385 ,orange1.1g006104m ,Eucgr.I01833.1 ,GSVIVT01032523001

Summary and Future Directions

The association of functional profiles with *cellulose synthase* superfamily phylogenetic structure is a central task in the understanding of cell wall carbohydrate biosynthesis and thus the evolution of the plant cell wall. The work described in this thesis aimed to reconstruct the evolutionary history of the *CesA* gene superfamily in the embryophytes using Bayesian and likelihood-based models to provide systematic structure and evolutionary inferences useful to future enzyme family functional characterisation. This chapter discusses the core findings of the thesis and outlines promising directions for future research.

The Poaceae (grasses) are a commercially significant monocot family and have attracted significant attention from the plant cell wall community, not least because they are unique in having an abundance of (1,3;1,4)-β-glucan in their cell walls. Whole genomes have now been sequenced for numerous grass species, providing a well sampled group in which to explore the evolutionary origins and dynamics of polysaccharide synthesis. In Chapter 2 we reconstructed the evolutionary history of the *CesA*, *CslD*, *CslE*, *CslF*, *CslH* and *CslJ* families in the Poaceae and investigated how selective forces have operated during evolution after ancestral gene duplications. We observed an asymmetric application of selection pressure across many lineages, with significant episodic selection following duplication events in *CesA*, *CslD* and *CslF* gene lineages. Together with a broad variation in nucleotide substitution rates, this indicated a dynamic history where gene duplicates have experienced varied and specific selection pressures. Additionally, we found sustained shifts in selection in the *CesA1*, *CesA8* and *CslF7* gene members, with each branch in the clade under positive selection. Whether such varied selection pressures reflect the synthesis of new polysaccharides, the organisation of protein complexes, deleterious mutation shielding, pathogen responses, or other forces, remains to be investigated. However, we were able to explore these changes in a structural context with the construction of a *CslF6* homology model, refined using molecular dynamics simulations. Here, we mapped the amino acid residues identified as being under positive selection onto the three-dimensional model and observed selection operating on: 1) the 'finger' helix that is responsible for translocating glucose units (Morgan et al., 2014), 2) residues within the substrate binding cleft, and 3) within the transmembrane pore from which the nascent polysaccharide emerges. By identifying such regions we provide targets for ongoing and future experimental work, and pose highly significant questions, such as whether the evolution of a particular dynamic associated with the *CslF6* finger helix contributed to the evolution of (1,3;1,4)-β-glucan biosynthesis. Finally, we observed that the primary cell wall associated *CesA* genes are nested within the secondary associated *CesA* clade. This finding suggests that the primary *CesA* gene family evolved from a secondary CesA ancestor.

In contrast to the clade comprising *CesA*, *CslD* and *CslF*, the *CslE*, *CslH* and *CslJ* clade was shown to have only one branch, leading to *CslE*, under episodic selection. This is consistent with the comparatively few ancestral duplication events in this group suggesting a less dynamic early evolutionary history. However, likelihood-based selection tests could not be performed on the *CslJ* clade because the level of historical divergence on the *CslJ* branch was too high for the evolutionary model used in these analyses. This finding was notable in that it implied either the loss of closer gene family relatives preceding the sampled species divergence, or represented a substantial shift in rates of sequence evolution.

High levels of divergence could indicate a major change in enzymatic function. Hence, in Chapter 3 we focused on the phylogenetic analysis and functional characterisation of the *CslJ* family. In this chapter we improved the taxonomic sampling of the *CslJ*, *CslE* and *CslH* family by including thirteen eudicot and two non-Poaceae monocot species, thus expanding previous analyses to include the *CslG* and *CslB* families. Our results yielded two lines of evidence that the Poaceae have lost ancient members of the *CslB*, *CslE*, *CslG*, *CslJ* clade. Firstly, the grass *Panicum virgatum* (switchgrass) and the non-grass monocot *Musa acuminata* (banana) were resolved to contain *CslG* members; this suggests that a wider population of Poaceae *CslB*, *CslE*, *CslG*, *CslH* and *CslJ* clade members were present in the common ancestor of grass and banana. Secondly, we found a much larger expansion of eudicot *CslJ* relatives than had been previously observed (Yin et al., 2009); these were present in a taxonomically broad sampling of eudicots, suggesting an early branching lineage that was lost in all non-grass monocots investigated here. *CslJ* family members were also shown to have high intra-clade sequence homology in comparison to their eudicot relatives, which we have named *CslM*, and were shown to have undergone a significant long term shift in selection pressure across eleven amino acid residues.

The possibility that the observed signal of positive selection on the *CslJ* clade reflects the acquisition of new protein function is intriguing. Indeed, the residues shown to be under selection since the Poaceae diverged from the eudicots (*CslJ* and *CslM* are reciprocally monophyletic) are distributed within the catalytic domain and some are within eight residues of the QxxRW motif. Chapter 3 contains experimental evidence that *CslJ* is implicated in the biosynthesis of (1,3;1,4)-β-glucan. It is possible therefore that the strong positive selection pressure operating on the *CslJ* family has been driven by the selective advantage of having three gene families (with *CslF* and *CslH*) involved in the production of (1,3;1,4)-β-glucan. An alternative scenario is that (1,3;1,4)-β-glucan synthase functionality was present in the *CslJ* family following the monocot-eudicot divergence, but was subsequently lost in all monocot lineages and some Poaceae, and selection targeted some other property such as tissue specificity, enzymatic rate, fine structure composition or pathogen response. Either way, the finding that three highly diverged gene families are implicated in (1,3;1,4)-β-glucan biosynthesis presents an intriguing case of convergent evolution and poses compelling questions relating to the evolutionary drivers and context of independent origins for (1,3;1,4)-β-glucan biosynthesis. This broad distribution of a particular polysaccharide across the *CesA* superfamily tree highlights the difficulty in mapping polysaccharide product to phylogenetic structure. While initial characterisation of cellulose- producing *CesA* genes and hemicellulose-producing *Csl* genes might remain valid, it is clear that the situation is more complex than initially assumed (Richmond and Somerville, 2000). Rigorous systematic analysis of the superfamily is therefore needed to extend previous studies of phylogenetic structure that used relatively few embryophyte species. Indeed the existing nomenclature could obscure a great deal of systematic variation, thus making the task of associating clade structure to polysaccharide product difficult. These observations and the confusion in the communication of *CesA* superfamily homologues among plant species prompted the development of a revised nomenclature system for angiosperm superfamily members, as presented in Chapter 4.

The phylogenetic analysis presented in Chapter 4 reveals a complex systematic structure. To encourage acceptance among the plant cell wall community, it was deemed important to retain the root gene family symbols. However, the gene trees reconstructed in Chapter 4 highlight the different evolutionary histories experienced by the monocots and eudicots following their divergence, with a number of monocot- and eudicot-specific families resolved within root clades. When this is considered alongside the presence of reciprocally monophyletic groupings such as *CslB*/*CslH* and *CslJ*/*CslM* that have now been proven to be functionally divergent, the possibility of further functional divergence within the established root families emerges. For instance, in the *CslC* clade, which has been implicated in xyloglucan backbone biosynthesis (Cocuron et al., 2007), we observe two large clades with a complex structure of gene loss following the monocot-eudicot divergence where monocot- and eudicot- specific clades are positioned as ancestral to large expansions of the alternate group. Considering xyloglucan is sparsely distributed across the monocots and relatively abundant among eudicots, the possibility of functional divergence between the two observed *CslC* clades complicates efforts to map polysaccharide synthase function to phylogenetic order. Despite these difficulties in systematic characterisation, the construction of a unified nomenclature system is a potentially important contribution that will facilitate communication when comparing homologues between species and provide a framework for testing targeted functional hypotheses. Newly discovered genes are often named according to model species such as Arabidopsis and rice (which do not share nomenclature) and usually in order of gene discovery. In Chapter 4 this problem is addressed by proposing a nomenclature that uses the phylogenetic branching order of ancestral gene duplications and characterises clades according to discriminatory protein motifs.

Future work

A fundamental limitation in phylogenetic studies is the trade-off between the genomic and taxonomic breadth of sampling. Because we used fully sequenced genomes in this study, in order to give a proper account of gene family presence and expansion, we were limited to sequencing project choices that often did not account for phylogenetic breadth but rather were based on other considerations such as the economic importance of the organism. One consequence of this is that monocot sampling is heavily skewed towards the agriculturally crucial grasses, with very few non-Poaceae sampled. This compares poorly with the relatively wide sampling of eudicot lineages. Future work could therefore be focused on adding species that expand taxonomic coverage particularly non-Poaceae monocots, non-agricultural grasses and species with unusual traits such as the small free living eukaryote, *Ostreococcus.* Non-Poaceae monocots will add resolution to the expansion of the various Poaceae- and eudicot-specific clades present within the superfamily and offer the possibility of discovering additional major superfamily divisions. Priority should also be given to the incorporation of early diverged plant lineages as they become available. The current very limited sampling of lower plants makes their inclusion in analyses problematic because it is difficult to determine how representative the available taxa are to their larger order. An appropriately sampled lower plant collection will be extremely valuable in determining very deep branching order of the family, for instance in resolving the placement of the *CslB* and *CslH* clade within the superfamily, and with the broader evolution of cell walls in general.

A significant goal of the plant cell wall community is a structural description of plant *CesA* superfamily members (Morgan et al., 2014; Sethaphong et al., 2013). Future work could benefit from a focus on leveraging structural observations with phylogenetic analyses. Ancestral state reconstruction can be used to predict the most likely sequences of specific nodes in the tree. These sequences thus represent the ancestral proteins of, for instance, the common ancestor of *CslJ* and *CslM* before the monocot-eudicot divergence. Such data can be used to inform functional characterisations of amino acid mutations in *CesA* superfamily enzymes because we can account for historical mutations and their effect on functional divergence. When comparing extant sequences, at the tips of the phylogenetic tree, it is difficult to determine whether mutations are relevant to functional divergence. By modelling amino acid substitutions through time we can predict residues that have little bearing on function. Additionally, ancestral state reconstruction provides an appropriate substitution background that can account for two classes of epistatic mutation: permissive mutations that tolerate other substitutions responsible for functional change, for example by increasing protein stability that permits the disruptive function-shifting mutation; and restrictive mutations that inhibit the function of certain enzyme family function and, if not accounted for, will mask observations of the contribution of specific mutations to functional divergence between closely related protein lineages (Harms and Thornton, 2010). Heterologous expression of ancestral *CesA* superfamily sequences, certainly a challenging prospect given that they are membrane-bound proteins, could nonetheless offer valuable insight into the functional evolution of this crucial plant cell wall biosynthesis enzyme family.

References

Cocuron, J.-C., Lerouxel, O., Drakakaki, G., Alonso, A. P., Liepman, A. H., Keegstra, K., Wilkerson, C. G. (2007). A gene from the cellulose synthase-like C family encodes a beta-1,4 glucan synthase. Proceedings of the National Academy of Sciences of the United States of America, 104(20), 8550–8555. http://doi.org/10.1073/pnas.0703133104

Harms, M.J and Thornton, J.W. Analyzing protein structure and function using ancestral gene reconstruction. Current Opinion in Structural Biology, 20(3), 360-366. doi:10.1016/j.sbi.2010.03.005.

Morgan, J. L. W., McNamara, J. T., & Zimmer, J. (2014). Mechanism of activation of bacterial cellulose synthase by cyclic di-GMP. Nature Structural & Molecular Biology, 21(5), 489–96. http://doi.org/10.1038/nsmb.2803

Richmond, T. a, & Somerville, C. R. (2000). The cellulose synthase superfamily. Plant Physiology, 124(2), 495–8.

Sethaphong, L., Haigler, C. H., Kubicki, J. D., Zimmer, J., Bonetta, D., DeBolt, S., & Yingling, Y. G. (2013). Tertiary model of a plant cellulose synthase. Proceedings of the National Academy of Sciences, 110(18), 7512-7517. http://doi.org/10.1073/pnas.1301027110

Yin, Y., Huang, J., & Xu, Y. (2009). The cellulose synthase superfamily in fully sequenced plants and algae. BMC Plant Biology, 9, 99. http://doi.org/10.1186/1471-2229-9-99

Appendix: Co-authored papers with contribution from this thesis

Schreiber, M., Wright, F., MacKenzie, K., Hedley, P. E., **Schwerdt, J. G**., Little, A., Burton, R.A., Fincher, G.B., Marshall, D., Waugh, R & Halpin, C. (2014). The barley genome sequence assembly reveals three additional members of the CslF (1, 3; 1, 4)-β-glucan synthase gene family. PloS one, 9(3), http://dx.doi.org/10.1371/journal.pone.0090888

Wong, S. C., Shirley, N. J., Little, A., Khoo, K. H., **Schwerdt, J.G.**, Fincher, G. B., Burton, R.A., & Mather, D. E. (2015). Differential expression of the HvCslF6 gene late in grain development may explain quantitative differences in (1, 3; 1, 4)-β-glucan concentration in barley. Molecular Breeding, 35(1), 1-12. http://dx.doi.org/10.1007/s11032-015-0208-6

Ermawar, R. A., Collins, H. M., Byrt, C. S., Betts, N. S., Henderson, M., Shirley, N. J., **Schwerdt, J.G**., Lahnstein, J., Fincher, G.B., & Burton, R. A. (2015). Distribution, structure and biosynthetic gene families of (1, 3; 1, 4)-β-glucan in Sorghum bicolor. Journal of integrative plant biology, 57(4), 429-445. http://dx.doi.org/10.1111/jipb.12338

Dimitroff, G., Little, A., Lahnstein, J., **Schwerdt, J. G.**, Srivastava, V., Bulone, V., Burton, R.A., & Fincher, G. B. (2016). (1, 3; 1, 4)-β-Glucan Biosynthesis by the CSLF6 Enzyme: Position and Flexibility of Catalytic Residues Influence Product Fine Structure. Biochemistry, 55(13), 2054-2061. http://dx.doi.org/10.1021/acs.biochem.5b01384

Ermawar, R. A., Collins, H. M., Byrt, C. S., Henderson, M., O'Donovan, L. A., Shirley, N. J., **Schwerdt, J.G**., Lahnstein, J., Fincher, G.B., & Burton, R. A. (2015). Genetics and physiology of cell wall polysaccharides in the model C 4 grass, Setaria viridis spp. BMC plant biology, 15(1), 236. http://dx.doi.org/10.1186/s12870-015-0624-0

Marcotuli, I., Houston, K., **Schwerdt, J. G.**, Waugh, R., Fincher, G. B., Burton, R. A., Blanco, A., & Gadaleta, A. (2016). Genetic Diversity and Genome Wide Association Study of β-Glucan Content in Tetraploid Wheat Grains. PloS one, 11(4), e0152590. http://dx.doi.org/10.1371/journal.pone.0152590