Assigned Reading:

Glycan Variation and Evolution in the Eukaryotes

Genome Analyses Highlight the Different Biological Roles of Cellulases
Glycan Variation and Evolution in the Eukaryotes
Glycan variation and evolution in the eukaryotes

Anthony P. Corfield¹ and Monica Berry²

¹ Mucin Research Group, University of Bristol, School of Clinical Sciences, Bristol Royal Infirmary, Bristol, BS2 8HW, UK
² University of Bristol, School of Physics, Centre for Nanoscience and Quantum Information, Tyndall Avenue, Bristol, BS8 1FD, UK

In this review, we document the evolution of common glycan structures in the eukaryotes, and illustrate the considerable variety of oligosaccharides existing in these organisms. We focus on the families of N- and O-glycans, glycosphingolipids, glycosaminoglycans, glycosylphosphatidylinositol (GPI) anchors, sialic acids (Sias), and cytoplasmic and nuclear glycans. We also outline similar and divergent aspects of the glycans during evolution within the groups, which include inter- and intraspecies differences, molecular mimicry, viral glycosylation adaptations, glycosyltransferase specificity relating to function, and the natural dynamism powering these events. Finally, we present an overview of the patterns of glycosylation found within the groups comprising the Eukaryota, namely the Deuterostomia, Fungi, Viridiplantae, Nematoda, and Arthropoda.

Glycan structure: why it matters

It is well known that the surface of cells at all levels of life is coated with an array of glycans, often termed the ‘glyocalyx’ [1–4]. The presence of this sugar coat is linked with biological functions that are fundamental to the normal behavior of the organisms throughout their life. Such functions range from imparting protein stability and function [5], to mediating the interaction of cells with the extracellular matrix [6–9], to sperm–egg binding processes [10], and host–pathogen interactions [11–16]. The biological choice of glycans as targets for such interactions reflects both their diversity and specificity. It also places demands on glycosylation to maintain and adapt to dynamic biological environments.

Here, we consider glycan structure and glycoconjugate function within the eukaryotes. Details for the individual phyla can be found in supplemental material online. A phylogenetic tree for eukaryotic species that have had their genomes sequenced is shown in Figure 1. The CAZy database (http://www.cazy.org) provides an invaluable source of information regarding all proteins that manipulate glycans in all their forms and is widely used as a primary target for genomic screening of species-specific glycomes [17]. The Consortium for Functional Glycomics website also provides a valuable tool for glycomic research (http://www.functionalglycomics.org). A consequence of the increased biological interest in glycans has been a focus on the chemistry–glyobiology frontier and the need to understand chemical and physical approaches for glycan analysis [3,4,18,19].

 Sugars chains face the outside world and can occupy most of the hydrodynamic volume of any molecules that carry them. Together with cognate molecules, they create bonds and assist intercellular recognition. Through spatial organization and concentration, they confer a further degree of selectivity to the formation of these bonds. Glycosylation is the default protein modification across prokaryotes [20] and eukaryotes; it affects fundamental and conserved aspects of the protein life cycle, including folding, intracellular trafficking, and co-translational quality control. The processes that enable protein maturation are homologous and occur in all kingdoms of life. Relevant to this discussion, the assembly of an oligosaccharide on the cytoplasmic side of the plasma membrane by a set of specific glycosyltransferases and the subsequent translocation of the lipid-linked oligosaccharide is a unifying scheme in all systems.

Changes to glycosylation in tumor cells [21] occur to both core and outer carbohydrate structures. Among the first are: (i) overexpression of β1–6 branching of N-glycan [22]; (ii) greatly enhanced presence of truncated T, Tn, and sialyl-Tn glycans on mucins [23,24]; and (iii) loss of GPI anchors [25]. Increased Sia content [26,27], and enrichment in hyaluronan in the stroma surrounding a tumor [28–30] are examples of peripheral glycan changes in malignancies.

Here, we focus on eukaryotic glycosylation, which represents only an instance of glycosylation, given that continuous glycan evolution is traceable through to Euacteria and Archaea. A more comprehensive review of bacterial glycosylation can be found in the article by Tan et al. in this special issue of TIBS [20]. Several earlier publications have addressed the variation of eukaryotic glycosylation compared with other organisms and their role in evolution (e.g., [31–38]). Furthermore, the scope of eukaryotic glycosylation is too broad to review in full detail; here, we highlight the fundamental patterns and concepts of glycan structure and evolution, illustrated with examples.
Glycan variation in the eukaryotes

Examination of glycoconjugate properties immediately uncovers a multiplicity of structures that require appraisal in terms of their function within these molecules. Table 1 shows examples of the major groups of proteins that manipulate glycoconjugates, highlighting the broad range of events that are involved in glycobiological metabolism.

Initial consideration of glycan structures must include the metabolic pathways that lead to their formation. The monosaccharides that form the building blocks are derived from diverse sources. Some are obtained from extracellular origins via salvage pathways from both external and intracellular sources, together with the major intracellular glycolytic pathways that feed monosaccharides into glycan synthesis and degradation.

A series of interlinked metabolic pathways produces the precursors of the nucleotide sugars, while sugar transporters facilitate the transport of glucose across the plasma membrane in either an energy-independent (facilitated diffusion transporters, the GLUT family), or an energy- and sodium-dependent way (SGLT family). Monosaccharides are activated to high-energy nucleotide sugars (UDP-Glc, UDP-Gal, UDP-GlcNAc, UDP-GalNAc, UDP-GlcA, UDP-Xyl, GDP-Man, GDP-Fuc, and CMP-NeuAc), which act as universal donors in glycosyltransferase reactions for each glycosynthetic pathway. Glycosulfotransferases utilize 3’-phosphoadenosinyl-5’-phosphosulfate (PAPS) as the active sulfate donor. Phosphorylated sugars, such as glucose-6-phosphate, mannose-6-phosphate, and glucosamine-6-phosphate, are similarly formed through the action of kinases with ATP as the phosphate donor.

Pathways leading to the nucleotide sugars are regulated through feedback inhibition of the key enzymatic steps initiating each route. The entry of glucose into the

| Table 1. Major groups of proteins manipulating carbohydrates in eukaryotes* |
|-----------------------------|----------------------------------|
| **Protein family** | **Main categories** |
| Glycosyltransferases | Transfer of mono- or oligosaccharides to acceptors, mono- and oligosaccharides, proteins, lipids, and DNA using nucleotide sugars and dolichol phosphate-mono and oligosaccharides |
| | A large family of enzymes catalyzing the transfer of fucose, mannose, glucose, galactose, xylose, N-acetylglucosamine, N-acetylgalactosamine, N-acetylmuramic acids, glucuronic acid, and Sias |
| | Also includes transfer of sulfate, phosphate, acetyl, methyl, pyruvate, and ethanolamine to glycans |
| Glycosidases | Catalyze the hydrolytic cleavage of the glycosidic bond in all glycoconjugates; function in biosynthetic and catabolic cellular processes |
| | Over 100 families identified |
| Polysaccharide lyases | Cleave polysaccharides containing uronic acid through a β-elimination mechanism forming a new reducing end at the cleavage site |
| | Nine families identified |
| Carbohydrate esterases | Remove acetyl groups from polysaccharides, including pectin, xylan, galactoglucomannan, rhamnogalacturan and glycan linked Sias |
| | 16 or more families identified |
| Carbohydrate-binding modules | Noncatalytic proteins, that bind soluble and crystalline carbohydrates; usually present in proteins together with catalytic domains in carbohydrate-active enzymes |
| | Found in glycoside hydrolases, polysaccharide lyases, polysaccharide oxidases, glycosyltransferases, and plant cell wall expansins |
| | 14 families identified |
| Glycan-binding protein (GBP) molecules | GBP’s enable binding of the carbohydrate-active enzyme to its substrate, thus increasing the local concentration of the enzyme and enhancing substrate degradation |
| | Examples of proteins containing GBPs: C-type lectins, proteoglycan core proteins, Type II membrane receptors, collectins, selectins, dectins, mannosic receptors, layilin, galectins, and siglec |
| Sugar transporters | Transport of sugars across the plasma membrane into cells |
| | Hexose transporters |
| | Energy-independent, facilitated diffusion, glucose transporters (GLUT protein family; SLC2 gene) in yeast and mammals |
| | Energy dependent; e.g., sodium-dependent glucose transporters (SLC5A gene) |

* A brief overview of eukaryote proteins that interact with carbohydrates. More detail can be found on the CAZY website (http://www.cazy.org), at http://www.cazypedia.org and on The Consortium for Functional Glycomics website (http://www.functionalglycomics.org). All three sites provide further links to other carbohydrate-relevant databases.

Figure 1. Eukaryote phylogenetic tree. Phylogenetic tree of the multicellular Eukaryota, showing the main groups discussed in this review. Reproduced from Essentials of Glycobiology, Chapter 19 page 282, with permission of Cold Spring Harbor Press, P. Gagneux and TD Pollard. Abbreviation: mya, million years ago.
pathways leading to the formation of the N-acetylhexosamines is regulated by glucosamine-6-phosphate synthase, which is feedback inhibited by the target nucleotide sugar UDP-GlcNAc [41,42]. This same nucleotide sugar is further metabolized to yield Sias and CMP-β-O-Sias. This pathway is subject to inhibition at the level of the initial enzyme UDP-GlcNAc epimerase/kinase [43]. Further integration of glycosylation with monosaccharide biology requires the presence of sugar transporters. The transport of monosaccharides and activated sugars as nucleotide phosphate or dolichol phosphate into the cell is a significant event, enabling optimal efficiency of glycosylation processes in general.

In addition to de novo biosynthetic routes, salvage pathways add monosaccharides released from glycoconjugates during normal cellular turnover. Most are released during lysosomal degradation. The impact of these pathways depends on the organ and tissue. The liver is a major site of glycoconjugate degradation and the recovery of monosaccharide products makes a significant contribution to the biosynthetic potential of new glycoconjugates. Integration of these pathways and their valuable monosaccharide products between tissues and organs constitutes a central facet of whole-body eukaryotic glycobiology. Examples of the variety of glycan structures found in eukaryotes are listed in Table 2.

### Molecular mimicry
Natural and synthetic mechanisms utilize glycan structures to mimic a variety of cellular structures. Pathogenic organisms synthesize antigens that are analogs of naturally occurring biotargets and, in this way, induce cellular responses that enhance the survival of the pathogen, but disadvantage the host [44]. Glycoengineering can be used in the same manner to create glycotopes that can be used therapeutically. Currently, there is keen interest in the application of glycoengineering in recombinant glycoprotein technology. Correct glycosylation for optimal biological relevance is paramount in such research [45–48].

### N-glycans
The N-glycan class represents a range of glycans found in β-N-glycosidic linkage of an N-acetyl-D-glucosamine residue with a peptide and/or protein asparagine. Targeting of N-glycans to appropriate asparagine residues in proteins is achieved through the recognition of a tri-amino acid sequon: asparagine-X-serine/threonine. X may be any other amino acid, barring proline. There are several common features of the N-glycans: first, the core comprises Manα1-6(Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAcβ1Asn-X-Ser/Thr and generates two antennae. Second, the core is extended to form three groups of N-glycan: oligomannose forms only contain mannos; complex type have sialylated N-acetylactosamine trisaccharides on each antenna and a fucose on the GlcNAc linked to asparagine; and hybrid types have a combination of oligomannose and complex units on the Manα1-6 and Manα1-3 antennae, respectively.

The process of protein N-glycosylation is conserved across eukaryotes. The precursor glycan is assembled by the stepwise glycosyl transfer of the individual sugars while attached to a lipid carrier (dolichol pyrophosphate)
The trimming and processing events occur in the ER and have many features that are conserved in eukaryotes. These link with crucial regulatory functions, notably the correct folding of the glycoprotein to ensure optimal biological activity. The folding mechanism is governed by the chaperones calnexin and calreticulin, which regulate the exit of glycoproteins from the ER. This cycle of events involves the monitoring of the glycoprotein after trimming of the triglucosyl unit by α-glucosidases I and II. If the glycoprotein is incompletely folded, an α-glucosyl-glycoprotein glucotransferase acts on the terminal α1-2 linked mannose on the α1-3 mannosyl antenna. The glycoprotein is then recycled through the calnexin/calreticulin complex to reassess the folding efficiency. Glycoproteins that fail to fold properly are eliminated after translocation to the ER cytoplasm, where the N-glycans are removed and the protein is degraded in a series of events termed ‘ER-associated degradation’ (ERAD) [49,51]. Correctly folded glycoproteins are further trimmed and glycosylated to generate the three main classes of N-glycan in addition to the mannos-6-phosphate-terminated N-glycans, which bind to the Man-6-P receptor and are delivered to endosomes and lysosomes. This series of events occurs in the Golgi apparatus. Distinct patterns of peripheral N-glycosylation are found in different phyla from plants and invertebrates to mammals [49,51,52].

O-glycans

O-glycosylation has also been termed ‘mucin-type O-glycosylation’. Mucins represent the major examples of glycoproteins that carry this type of glycan. Glycans are attached to the protein backbone through α-linkage of N-acetyl-D-galactosamine to the side-arm hydroxy groups of serine and threonine residues. The addition of the GalNAc residue to generate the initial residue of mucin-type glycans, in addition to proteins that carry single GalNAc moieties, is catalyzed by a family of GalNAc transferases that has been extensively researched by the group of H. Clausen [53]. The mucins are a family of glycoproteins having serine-threonine-proline-rich variable number tandem repeat (VNTR) domains, which are the sites of multiple O-glycan substitution. They are found commonly at mucosal surfaces in mammals, where they have a protective role [1]. The O-glycosylation patterns found in the different mucins have formed the basis for review of the range of O-glycan structures and O-glycosylation pathways [1,40,54]. In addition to mucins, many other glycoproteins have peptide domains with mucin-like properties, but without the VNTR patterns that define the mucins themselves. The glycosylation of these domains follows the same biosynthetic routes reported for the mucins.

The mucins have a series of core structures that form a base for extended glycans. These show linear Galβ1-3GalNAc- (core 1) and GlcNAcβ1-3GalNAc- (core 3) structures and branched Galβ1-3(GlcNAcβ1-6)GalNAc- (core 2) and GlcNAcβ1-3(GlcNAcβ1-6)GalNAc- (core 4) structures. Extension may occur via the addition of backbone repeat units of N-acetyllactosamine types I and 2 (Galβ1-3GlcNAc- and Galβ1-4GlcNAc-, respectively), or as the branched I and

on the cytoplasmic face of the endoplasmic reticulum (ER). Addition of the first sugar, GlcNAc, to dolichyl phosphate involves transfer of a phosphate group and the sugar to form dolichyl pyrophosphate-GlcNAc. This product is extended to yield dolichyl pyrophosphate-GlcNAc2Man5 before being translocated to the luminal surface of the ER membrane by a flippase. Here, further mannose residues are transferred and the Mano1-3 antenna is elongated with two mannoses terminated by a triglucosyl unit Glc3a1-2Glc1a1-3Glc1-3-.. These glycosylation steps use dolichol-P-mannose and dolichol-P-glucose as donors rather than UDP-Glc and GDP-Man, the corresponding nucleotide sugar donors, which are substrates for the glycoprotein transfer reactions on the cytoplasmic side of the ER. The final dolichol-linked glycan, Glc5Man5GlcNAc2, is then recognized by an oligosaccharyltransferase in the ER lumen that mediates attachment to protein asparagines to yield the β-N-glycosidic linkage [49,50]. This glycan serves as a substrate for trimming and processing reactions to yield the range of N-glycans found in nature. The biosynthetic pathway is found in all metazoans, plants, and fungi.
linear i antigens (Galβ1-4GlcNAcβ1-3Galβ1-4-). Subsequent peripheral substitution leads to complex oligosaccharides with ABO and Lewis blood group structures and a range of sialylated, fucosylated, and sulfated glycans [1,40,54]. The biosynthetic pathways leading to the formation of the mucin type O-glycans are well known, and an example is shown in Figure 2.

A large group of nuclear and cytoplasmic proteins carry a single sugar, O-β-N-acetyl-D-galactosamine (O-β-GlcNAC), attached to serine or threonine hydroxyl residues [55,56]. Most of these proteins have multiple O-β-GlcNAc additions at different sites. The common serine or threonine sites identified for O-β-GlcNAc substitution are also common sites for O-phosphorylation. Alternate O-β-GlcNAcylation and phosphorylation has been identified for several nuclear and cytoplasmic proteins. This modification is never extended and, furthermore, is a dynamic glycosylation event, with cyclic patterns mediated by two specific enzymes, an O-GlcNAc transferase (OGT) and an N-acetyl-D-glucosaminidase (OGlcNAcase). The latter recognizes and removes the protein-bound O-GlcNAc. This post-translational modification is abundant and common to all metazoans [55,56].

Several other O-glycans exist and, although these may not be as abundant or widely distributed as the mucins, they all have significant biological relevance.

O-Mannosylation comprises a range of glycans with mannose alpha linked to serine or threonine found in metazoans [57,58]. Major sites of expression are skeletal muscle and the nervous system, particularly the brain. Most structures reported relate to Neu5Aca2-3Galβ1-4GlcNAcβ1-2ManoSer/Thr, which is the major glycan in α-dystroglycan, a skeletal muscle glycoprotein. However, other glycans have been found, including fucosylated and glucuronic acid-3-sulfated forms and branched chains [57]. The Drosophila melanogaster O-glycan repertoire contains two genes encoding O-mannosyltransferase (POMT1 and POMT2), which catalyze this reaction. O-β-glucose and O-α-fucose are both found in the epidermal growth factor (EGF)-like repeats of Notch and Cripto/FRL/Critic proteins. These monosaccharides are linked to serine or threonine.

O-fucosylation shows a consensus sequence C3X3N/C(Ser/Thr)C3, with C2 and C3 being the second and third conserved cysteines in the EGF-like repeats. Two specific protein-O-fucosyltransferases (POFUT-I and POFUT-II) are responsible for this transfer, with GDP fucose as the donor substrate [59]. POFUT-I catalyzes transfer to EGF domains, while POFUT-II shows specificity for thrombospondin type-1 repeats [60]. The largest structure found is a tetrasaccharide, Neu5Aca2-3Galβ1-4GlcNAcβ1-3Fucα- O-Ser/Thr. Urokinase, factor XII, Cripto factor IX, Thrombospondin type-1 repeats, Notch, Delta, and Serrate, among others, have EGF-like repeats with both single O-fucose and tetrasaccharide glycosylation [59,60].

O-β-galactosylation occurs at serine and threonine residues in the EGF-like repeats, with the consensus sequence C3X3Ser/XPC2 distinct from the O-fucosylation sites. The most common glycan identified is a trisaccharide with two xylose residues (Xylo1-3Xylo1-3Glcβ-O-). Some EGF-like repeats also have single O-β-glucose additions, whereas few proteins show this modification: only factor VII, factor IX, and Notch [61,62].

O-β-galactosylation of collagen has been known for some time. This modification occurs together with the creation of hydroxylysine and hydroxyproline in the collagen protein and the addition of a disaccharide Glcα1-2Galβ-O-Hyl/Hyp [63].

C-mannosylation
This modification entails the attachment of a single mannose residue, in a C-linkage, to the indole ring of tryptophan. It is found in most eukaryotes, except yeasts, and is absent in bacteria. A peptide consensus sequence, W-X-X-W, carries the mannose residue on the initial tryptophan [64,65]. It also occurs in the CYS domains of MUC5AC and MUC5B [65]. The biosynthesis involves dolichol-phosphate-mannose as the donor [64]. C-mannosylation is thought to have a role in protein folding.

GPI anchors
GPI anchors are a common means of tethering proteins, through their carboxyl terminal, to the outer leaf of the cell membrane lipid bilayer, and presenting them to the external environment. GPI anchors are found throughout eukaryotes, including protozoa, plants, fungi, invertebrates, and mammals [66]. The membrane-mediated biosynthesis of GPI anchors is complex, but proceeds through conventional pathways and involves ten steps to yield a membrane GPI anchor-linked protein. A final step cleaves the inositol-bound palmitate from the anchor and enables phosphatidylinositol phospholipase C (PI-PLC) to release the protein from the membrane [66].

The general structure of the GPI anchors includes a common core comprising ethanolamine-phosphate-6Mano1-2Mano1-6Mano1-4GlcNa1-6N-acetyl-1-phosphatidyl-lysine. Proteins are attached to the amino group of the ethanolamine through their C-terminal carboxyl groups.
Table 3. Variety of Sias in eukaryotes

<table>
<thead>
<tr>
<th>Sia feature</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>Neu (a)</td>
<td></td>
</tr>
<tr>
<td>Neu5Ac (c)</td>
<td></td>
</tr>
<tr>
<td>Neu5Gc (d)</td>
<td></td>
</tr>
<tr>
<td>Kdn (b)</td>
<td></td>
</tr>
<tr>
<td>Chemical modifications</td>
<td>Mono-O-acetylation at positions 4, 7, 8, and 9; di-O-acetyl at 4, 9, 7, 8, and 8,9; and possibly tri-O-acetyl form at 7,8,9.</td>
</tr>
<tr>
<td></td>
<td>O-sulfate at position 8</td>
</tr>
<tr>
<td></td>
<td>O-methyl at position 8</td>
</tr>
<tr>
<td></td>
<td>O-lactyl at position 9</td>
</tr>
<tr>
<td></td>
<td>O-phosphate at position 9</td>
</tr>
<tr>
<td></td>
<td>Dehydro forms, with double bond in ring at position 2-3</td>
</tr>
<tr>
<td>Glycosidic linkage</td>
<td>α2-3</td>
</tr>
<tr>
<td>Adjacent sugars as a carrying glycan</td>
<td>Galactose N-acetyl-galactosamine</td>
</tr>
<tr>
<td>Glycan type</td>
<td>N-linked; O-linked; glycosphingolipid; and polysaccharide</td>
</tr>
<tr>
<td>Cellular organisation</td>
<td>Cell surface plasma membranes; glycolcalyx; glycosynapase</td>
</tr>
</tbody>
</table>

Different GPI anchors arise through substitutions to this core structure. Commonly found additions are the attachment of: (i) alkyl or alkenyl groups to the lipid terminus; (ii) palmitate to the C2 of myo-inositol; (iii) ethanolamine phosphate on C2 of Mano1-4 or C6 of Mano1-6; or (iv) glycans on each of the core mannoses. GPI-tethered proteins may be released from the plasma membrane by the action of PI-PLC: this property has significant biological relevance in cell surface protein expression. The substitution of palmitate on the C2 of myo-inositol blocks the action of PI-PLC and provides a significant regulatory modification. Many proteins with GPI anchors have been identified and the biological impact of this modification is evident in examples where both GPI-anchor tethered and soluble forms exist, relating to tissue distribution or developmental programmes; see [66] for a summary of examples.

Sias

The Sias represent a family of over 50 derivatives with unique diversity, wide occurrence, and miscellaneous biological functions [67]. They are not found in plants, prokaryotes, arthropods, and most invertebrates. An overview of the molecular and biological features of the Sias is given in Table 3.

The Echinodermata are the earliest clade identified to have Sias of various types [67,68]. The mammals express a plethora of Sias whose biology has attracted a large body of research.

Sia is a nine-carbon monosaccharide, comprising a six-membered ring with a three-carbon C7-C9 glycerol chain attached at ring carbon 6. Chemically, this is 5-acetamido-2-keto-3,5-dideoxy-D-glycero-D-galactononic acid; an unwieldy and impractical name that is effectively shortened to Sia, and abbreviated as ‘Neu5Ac’ or ‘Sia’ as a general form. A carboxyl group is attached at ring carbon 2 and an amino group is substituted on ring carbon 5. The amino group exists as acetyl (Neu5Ac) and glycolyl (Neu5Gc) forms. These two Sias are parent forms and both accept all of the modifications found in this family of monosaccharides. In brief, O-acetylation may occur at positions 4, 7, 8, and 9; O-methyl and O-sulfate at position 8, and O-lactyl, phosphate, and O-sulfate at position 9. The glycosidic hydroxyl at position 2 is in the α or β configuration for the free monosaccharide and forms the basis of glycosidic linkage to other sugars. Thus, α2-3, α2-6, and α2-8 (to other Sias) are common structures. Monosaccharide anhydro forms also exist, where a double bond is formed between C2 and C3, yielding, for example, Neu2en5Ac [67].

Biosynthesis of the Sias is initiated de novo from UDP-GlcNAc by the action of the UDP-GlcNAc epimerase/kinase complex and culminates in the formation of CMP-β-O-Neu5Ac. The de novo pathway also connects with salvage pathways, where Neu5Gc may be derived from the diet, and progresses through the activation mechanisms to generate CMP-β-O-Neu5Gc. The CMP-β-O-Sias are the donors in the sialyltransferase reactions leading to the formation of the library of sialglycoconjugates found in nature. In line with the broad range of potential acceptors, the sialyltransferase family mediates the biosynthesis of glycoconjugates in the Golgi. Some eukaryotic sialyltransferrases exhibit substrate specificity towards different

The Sias exist in nature as terminal sugars in many glycans and also in monosaccharide form as their dehydro analogs. A range of molecular characteristics leads to the abundance and variety of Sias. Key: yellow circles, D-galactose; yellow squares, N-acetyl-D-galactosamine; blue circles, D-glucose; blue squares, N-acetyl-D-glucosamine; blue/white squares, D-glucosamine; green circles, D-mannose; red triangles, L-fucose; purple diamonds, N-acetyl-D-neuraminic acid (Neu5Ac); light blue diamonds, N-glycolyl-D-neuraminic acid (Neu5Gc); blue/white diamonds D-glucuronic acid; orange/white diamonds, L-iduronic acid; orange stars, D-xylene; white diamonds, myo-inositol.
acceptors, glycoproteins, gangliosides (glycolipids), and polysaccharides, including polysialic acid chains. These transferases contain peptide sequences termed ‘sialylmotifs’ and are found in most eukaryote enzymes, apart from those in bacteria, suggesting different evolutionary routes [69]. Several Sia-related monosaccharides, such as 2-keto-3-deoxynonic acid (Kdn), are found at lower levels in most eukaryotes. Transfer of these analogs is also catalyzed effectively by the sialyltransferases. A novel form of sialyltransfer has been identified in the protozoa: trypanosomes have a trans-sialidase, which recognizes cell surface-bound Sias, cleaves them, and transfers them to their own cell surface acceptors to evade host immune screening [70].

The presence of a CMP-β-O-sialate transporter in the Golgi apparatus, but not in the ER, confirms the specificity of subcellular sialylation in eukaryotic cells [39].

In nonhuman species, the conversion of Neu5Ac to Neu5Ge is catalyzed by a specific CMP-β-O-Neu5Ac hydroxylase, encoded by the CMAH gene, which has been inactivated in humans [67]. O-acetylation of the 4-O and C7-C9-O positions is carried out enzymatically. Evidence for enzymatic O-acetyl transfer to a single position with subsequent migration of the acetyl esters to the nonoccupied positions has been proposed [67]. Enzymatic transfer of methyl groups by S-adenosylmethionine and sulfate by PAPS has been confirmed [67], but the mechanism of lactosylation has not been established.

Sialidases (neuraminidases) that release α-linked Sias from sialoglycoconjugates are known as NEU enzymes (NEU1–4). These show specificity for the type of sialoglycoconjugate and also have a characteristic tissue and cell expression pattern. By contrast, other eukaryotes, such as the fungi and invertebrates, have sialidases with broader substrate specificities, reflecting the strategy required to manipulate the sialoglycoconjugates encountered in their environment. Transport of Sias out of the lysosomes to the cytoplasm requires the action of sialin, a Sia exporter, but no plasma membrane-located Sia transporter has been identified so far.

The only Sia found in beta-linkage is CMPβ-O-Sia. This sialoside can be cleaved enzymatically by CMP-β-O-sialate hydrolase, a β-sialidase. Although this enzyme has been detected, there is only limited information concerning such activity and this remains an area for further study relative to the regulation of the sialone. The removal of O-acetyl esters is governed by sialate O-acetyl esterases, which have selective or broad specificities for the site of O-acetylation. Release of 4-O-acetyl, but not C7-C9, or mono 9-O-acetyl, but not 7,9-di-O-acetyl implies roles for different sialate O-acetylation patterns.

In accord with the range of sialoglycoconjugates and their sialylation patterns, Sia-binding lectins have been identified from many sources and are features of glyco-biology throughout the eukaryotes. The vertebrates have two main classes of lectin: (i) the C-type lectins, the selectins; and (ii) the I-type lectins, the siglecs. These proteins have important roles in cell–cell recognition processes. Accordingly, the glycans that carry the Sias have a role in recognition. Furthermore, the various modifications of the Sias, especially O-acetylation, may mediate the interactions between cells [67]. The specific binding properties of these lectins have led to their use as tools in monitoring Sia expression and events facilitated by Sias [71–73].

Glycosphingolipids and glycosaminoglycans are described in other reviews in this Special Issue and so are not discussed here.

**Evolutionary variation of glycans**

This final section emphasizes some of the main issues that relate to the study of glycans. These are essentially general comments drawn from the glyobiological literature and relate to evolutionary aspects of this work.

**General diversity**

Monosaccharides are particularly well suited to construct a range of different sequences, due to their ring and open chain patterns with equatorial and axial arrangement of the hydroxyl groups. In addition, the anomeric hydroxyl group enables the formation of alpha and beta glycosides for each monosaccharide. Thus, when the complete range of glycan structures found in the eukaryotes is screened, the initial impression is one of enormous scope and diversity. Closer examination shows that, although many monosaccharides exist chemically, only a relatively small selection of the total has been adopted to build biological sequences, and this reflects evolutionary selection.

It is striking how many glycan structures are shared by different groups of eukaryote. This can be explained by the evolution of proteins that control glycan metabolism. The glycosyltransferases, glycosyl hydrolases, sugar transporters, and lectins are organized in each tissue and cell to deliver the required glycosylation of the glycoconjugates, providing the structural and functional needs in each case. All of the families of glycans are represented in eukaryotes: N-glycans, the different O-glycans, C-mannose, glycolipids, and glycosaminoglycans. Each glycan type exhibits a core unit that is common to all. Extension of these core units also follows shared processes. However, completion of the glycans through peripheral additions results in strain- and phylum-specific structures. These are the moieties that face the outer world.

**Natural forces powering glycan evolution**

The glycan diversity found in individual environments is formed through known metabolic pathways. Therefore, many features of glycan structure influence phenotypic variation and serve to emphasize the flexibility of utilizing sugar-based sequences [18,23,37,38,51,54,74–78]. The glycobiological process is powerful because it enables a variety of possibilities to be selected in exclusive and appropriate conditions. Evolutionary analysis shows that unicellular organisms create glycans from a range of monosaccharides, while the eukaryotes use a smaller number [79]. Structural selection occurs by mutation or through exclusion pathways [80]. The basis of glycan structure and recognition at chemical and biochemical levels is robust. Glycoforms are generated through the action of the glycosyltransferases and degraded by the glycosidases. The pattern of expression of these enzymes, combined with their spatial and temporal properties, leads to a predictable range of glycan
structures, which can be matched with those detected by chemical methods in vivo. A further level of recognition arises due to spatial clustering and ligand density.

Concluding remarks
Glycans are generated through the action of glycosyltransferases and degraded by glycosidases. The pattern of expression of these enzymes, combined with their spatial and temporal properties, leads to a predictable and dynamic range of glycans structures, which share common features of monosaccharide composition, sequence, and conformation, and also confer cell, tissue, and clade specificity. Recognition relies on not only ligand presence, but also ligand density and its spatial organization; that is, clustering. These molecular patterns form a basis for the recognition of self and non-self.

Acknowledgments
We wish to thank Hans-Joachim Gabius for his guidance and encouragement during the preparation of this review. We are also grateful to Pascal Gagneux and P.T. Pollard for permission to reproduce the phylogenetic tree in Figure 1. Due to editorial constraints, it is not possible to cite all of the relevant literature linked with the broad topic of eukaryote glycosylation. Therefore, we acknowledge the contributions made by authors whose work it was not possible to reference here.

Appendix A. Supplementary data
Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.tibs.2015.04.004.

References
Review


45 Castillo, A. et al. (2013) Generation of biologically active multimodified recombinant human EPOFs in plants. *PLoS ONE* 8, e54836


78 Varke, A. (1993) Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* 3, 97–130


Genome Analyses Highlight the Different Biological Roles of Cellulases
Genome analyses highlight the different biological roles of cellulases

Felix Mba Medie1,2, Gideon J. Davies3, Michel Drancourt1 and Bernard Henriussat2

Abstract | Cellulolytic enzymes have been the subject of renewed interest owing to their potential role in the conversion of plant lignocellulose to sustainable biofuels. An analysis of ~1,500 complete bacterial genomes, presented here, reveals that ~40% of the genomes of sequenced bacteria encode at least one cellulase gene. Most of the bacteria that encode cellulases are soil and marine saprophytes, many of which encode a range of enzymes for cellulose hydrolysis and also for the breakdown of the other constituents of plant cell walls (hemicelluloses and pectins). Intriguingly, cellulases are present in organisms that are usually considered as non-saprophytic, such as Mycobacterium tuberculosis, Legionella pneumophila, Yersinia pestis and even Escherichia coli. We also discuss newly emerging roles of cellulases in such non-saprophytic organisms.

Cellulose is the most abundant form of photosynthetically fixed carbon in the biosphere. It is a fibrous polymer of glucose units that are linked by β-1,4-glycosidic bonds, and it occurs naturally in plants and in several other organisms, including certain bacteria, some fungi, protozoa (such as Naegleria gruberi1, Dictyostelium discoideum2 and Acanthamoeba castellanii3) and in one group of animals, the urochordates or ‘tunicates’ (REF. 4). Natural cellulose chains are aggregated laterally by means of hydrogen bonding and van der Waals’ forces to generate microfibrils of parallel chains5 that crystallize as two forms, Iα and Iβ6. Plant cellulose is further surrounded by a network of other polymers, including hemicelluloses, pectins and lignin7. Hemicelluloses are branched polysaccharides with a backbone of neutral sugars that form hydrogen bonds to the surface of cellulose fibrils, whereas the major component of pectins is galacturonic acid. The simplest pectin is homogalacturonan, an unbranched polymer of α-1,4-d-galacturonic acid8, but some pectins can exhibit considerable chemical complexity, such as the rhamnogalacturonans, which display numerous and varied carbohydrate side chains. Lignin fills the space between the cellulose, hemicellulose and pectin in the plant cell wall; it consists of a highly crosslinked matrix of aromatic residues.

The capacity to convert cellulose in complex plant cell wall substrates is of crucial importance for the carbon cycle9, reflecting the abundance of cellulose in nature. However, cellulose is notoriously difficult to hydrolyse enzymatically because it contains resilient glycosidic bonds10, is crystalline and is tightly associated with other polysaccharides (as mentioned above). The complete digestion of cellulose is therefore largely restricted to a specific group of cellulolytic microorganisms that produce complex combinations of enzymes (cellulases, hemicellulases and pectinases) which act synergistically to break down cellulose and its associated cell wall components. These enzymes belong to different sequence-based families of glycoside hydrolases (GHs) — which include cellulases that cleave the β(1→4) bonds of cellulose, making them attractive candidates for the sustainable production of cellulosic ethanol from lignocellulosic materials11–13 — carbohydrate-binding modules (CBMs), polysaccharide lyases and carbohydrate esterases, as classified in the Carbohydrate-Active Enzyme database (CAZy)14. Cellulolytic bacteria have traditionally been searched for in a wide variety of environmental and industrial niches, including soil, marine sediments and, more recently, the gut microbiota of animals and humans. Currently, the prevailing dogma is that cellulolytic bacteria live a saprophytic lifestyle.

However, the view that cellulases act only in the domain of plant cell wall degradation has recently been challenged by the demonstration that Mycobacterium tuberculosis, one of the pathogens responsible for animal and human tuberculosis, has three active cellulose-targeting proteins, encoded by Rv0062 (also known as celA1 (REF. 15); belonging to family GH6), Rv1090 (also known as celA2b; belonging to family GH12) and Rv1987 (REFS 16,17) (belonging to family CBM2). Similarly, the pathogen Legionella pneumophila secretes an endoglucanase18 (belonging to family GH5). These

© 2012 Macmillan Publishers Limited. All rights reserved
observations demonstrate that the analysis of complete genome sequences can reveal previously unconsidered metabolic traits and that the prevailing view that cellulases are restricted to saprophytic organisms is incorrect. It is becoming clear that cellulases are also involved in cellulose biosynthesis, as originally described for the bacterium *Gluconacetobacter xylinus* (formerly known as *Acetobacter xylinum*) and perhaps best characterized subsequently for the plant protein KORRIGAN, which is required for the assembly of cellulose and hemicellulose networks in *Arabidopsis thaliana*. In this Analysis article, we review some of the more unusual and indeed counter-intuitive roles for cellulases that are now emerging from wide-scale analyses of bacterial genomes.

**Cellulose and its enzymatic breakdown**

Two main types of cellulase enzyme activity are well characterized. Endoglucanases (ENZYME entry EC 3.2.1.4) hydrolyse internal bonds at random positions of the less ordered (or amorphous) regions of cellulose. These enzymes generate chain ends for the processive action of the second type of cellulases, the cellobiohydrolases (which are exoglucanases). These act in a unidirectional manner from either the non-reducing (ENZYME entry EC 3.2.1.91) or the reducing (ENZYME entry EC 3.2.1.176) ends of cellulose polysaccharide chains, liberating cellobiose as the major product. β-d-glucosidases (ENZYME entry EC 3.2.1.21) further hydrolyse cellobiose and thus relieve the system from end product inhibition. Various types of synergies have been observed between different sorts of cellulase during the breakdown of crystalline cellulose, but convincing explanations at the molecular level have yet to be found, especially for the cooperation between cellobiohydrolases. The cooperation between endo- and exo-acting cellulases is intuitively easier to understand, but the degree of synergy observed is usually less than that expected for a system in which the cellobiohydrolase depends entirely on the endo-glucanase for the provision of free chain ends. In an illustration of the poor understanding of the role of cellobiohydrolases, a recent investigation has shown that the two major cellobiohydrolases of *Clostridium thermocellum* can be deleted without severely affecting cellulose breakdown.

Other enzymes besides the endo-acting and exo-acting cellulases participate in the breakdown of cellulose. Some of these enzymes use oxidative chemistry, as occurs in the cleavage of chitin (a fibrous, crystalline polymer of β-1,4-N-acetylglucosamine residues) by the bacterial enzyme chitin-binding protein 21 (Cbp21); this mode opens up the crystalline polysaccharide material that is inaccessible for hydrolysis by other GHs. Furthermore, the reductant ascorbic acid boosts the efficiency of Cbp21 (REF. 23). Recently, a *Streptomycetes coelicolor* homologue of Cbp21 was shown to act synergistically with cellulases in the digestion of cellulose. Similarly, GH61 enzymes, which have structural and active centre similarity to Cbp21 and its homologues, also facilitate the degradation of cellulose through an
CBM2

Acidothermus cellulolyticus (AKB52390)
- CBM3 - GH48 - FN3 - CBM2

Acidothermus cellulolyticus (AKB52388)
- GH6 - CBM3 - FN3 - GH12 - CBM42

Thermotoga maritima (AAD35393)
- GH74

Caldicellulosiruptor bescii (ACM60948)
- GH74 - CBM3 - CBM2 - GH48

Mycobacterium vanbaalenii (ABM14300)
- CalX - CalX - GH18 - CalX - CBM4 - CBM45 - CalX - CBM45 - CalX - GH5 - CalX

Clostridium thermocellum (AAA23225.1)
- GH26 - GH5 - CBM13 - DOG1

Caldicellulosiruptor saccharolyticus (ABP66297)
- GH5 - CBM28 - SLH - SLH - SLH

Amycolatopsis mediterranei (ADJ49127)
- CE3 - GH5 - FN3 - CBM2

Figure 2 | Modularity of cellulosytic enzymes. Shown are some examples of bacterial cellulases that contain different domains. The Protein database accessions are given in brackets for each protein. CalX, Calx-β motif; CBM, carbohydrate-binding module; CE3, carbohydrate esterase family 3; DOG1, dockerin; EXPN, expansin; FN3, fibronectin type III; GH, glycoside hydrolase; SLH, surface layer homology.

© 2012 Macmillan Publishers Limited. All rights reserved
Most bacteria that encode a cellulase have only one cellulase gene in their genome; such organisms were present in all phyla analysed. Bacteria with two or three cellulases genes were present in all phyla except for Synergistetes, Cyanobacteria and Deinococcus–Thermus (phyla for which all bacteria have only one or no cellulase) (FIG. 3b). Interestingly, few bacteria have more than three cellulases (FIG. 3b); these bacteria are mostly found in the phyla Actinobacteria, Bacteroidetes, Thermotogae and Chloroflexi and in the class Clostridia, and also include some members of the classes Gammaproteobacteria, Alphaproteobacteria, Bacilli and Betaproteobacteria (FIG. 3b). Bacteria that digest plant cell walls often encode multiple cellulases, together with hemicellulolytic and pectinolytic enzymes (TABLE 1). Thus, the co-occurrence of genes encoding cellulolytic enzymes and those encoding hemicellulolytic and/or pectinolytic enzymes accompanies a true saprophytic lifestyle.

**Saprophytic bacteria.** The saprophytic bacteria can be divided into two groups. The first group, referred to here as cellulase- and hemicellulase-containing saprophytes, was defined on the basis of the presence of at least one cellulase gene and three or more hemicellulases or pectinases. The 287 bacteria in this group (see Supplementary information S2 (table)) represent about half of the bacteria that encode cellulases. Of these, 92 can also produce cellulose (see below). The second group consists of 42 bacteria and is referred to as hemicellulose-containing saprophytes or non-cellulolytic saprophytes, as the genomes contain hemicellulases and pectinases but lack cellulase (see Supplementary information S3 (table)). This group includes human gut bacteria such as *Bacteroides thetaiotaomicron*, *Bacteroides vulgatus*, *Parabacteroides distasonis* and *Faecalibacterium prausnitzii* str. SL3/3. The group also contains members of the genus *Yersinia*, which includes major pathogens such as the plague agent, *Yersinia pestis*. All the *Yersinia* spp. genomes analysed contained cellulases and pectinases and were consequently assigned to the ‘cellulase- and hemicellulase-containing saprophytes’ category (see Supplementary information S2 (table)). This suggests that these bacteria have hitherto undescribed contact with plant material.

**Cellulose-synthesizing bacteria.** Approximately 11% (168 taxa) of the ~1,500 bacteria analysed are cellulose-synthesizing non-saprophytes (FIG. 5a). Numerous bacteria produce cellulose as a biofilm matrix polymer using the enzymes encoded by the bacterial cellulose synthesis (*bcs*) operon. For example, in *G. xylinus* the cellulose synthesis pathway is encoded by the *bcsA*, *bcsB*, *bcsC* and *bcsD* genes, which together form an operon that is conserved in other cellulose-producing bacteria such as *Escherichia coli* and *Salmonella enterica* subsp. *enterica* serovar Typhimurium. Interestingly, the *bcs* operon
also encodes a cellulase of the GH8 family, BcsZ, the exact function of which in cellulose biosynthesis remains unclear. It was thought initially that BcsZ acted as a transglycosidase rather than an endoglucanase, but such a role is unlikely because transglycosylation is catalyzed by retaining glycosidases, whereas GH8 family enzymes use the inverting mechanism (see BOX 1). The processive glycosyltransferase 2 (GT2) family (members of which are called BcsA, CesA or AcsA in various species) has since been recognized as the enzyme family that polymerizes cellulose directly from UDP-glucose; the role of BcsZ in cellulose synthesis was confined (with no direct evidence) to nicking hypothetical defects or twists in the cellulose chain before cellulose crystallization into a microfibril. Another role for the cellulase could be to promote detachment or escape of the bacterium from cellulose that it has synthesized. However, this detachment or escape hypothesis does not explain the loss of cellulose biosynthesis capacity on addition of antibodies against G. xylinus BcsZ. In addition, genes involved in cellulose synthesis and those important for escape are more likely to be the subject of different regulatory elements (as opposed to being under co-regulation in an operon), as the two pathways are probably active under different environmental conditions.

Genomic analyses of bacteria show that cellulose biosynthesis operons are widespread and that, besides the processive GT2 family gene, the presence of a GH-encoding gene is a defining feature. Although the absence of operon structures in eukaryotes masks the direct partners of the processive GT2 protein in these organisms, in several cases a cellulase domain is found to be fused to the processive GT2 domain, indicating not only that the two enzymes are implicated in the same process, but also that the stoichiometry of the two proteins is 1/1. Such a stoichiometry argues for an important, but as-yet-uncharacterized, role for the hydrolytic component in polysaccharide biosynthesis.

Of the bacteria encoding cellulases that are likely to be involved in cellulose synthesis, 168 are non-saprophytes (see Supplementary information S4 (table)), whereas 92 are saprophytes, indicating that cellulose biosynthesis and saprophytic behaviour are not mutually exclusive. Furthermore, polymerizing systems that are analogous to the bcs operon, featuring a processive GT2 protein in tandem with a GH protein, may exist in bacteria and in eukaryotes. Instead of the family GH8 cellulase found in bcs operons, other GH families from the CAZy database were identified in these alternative systems, including GH5, GH6, GH12, GH18 and GH26, suggesting that convergent evolution has occurred. Strikingly, all of these hydrolytic families act on β-linked polysaccharides, raising the possibility that the respective operons or fusion proteins synthesize β-linked polysaccharides other than cellulose.

The role of the cellulases and hydollases in these alternative polysaccharide biosynthesis machineries remains unknown, but the recent finding that peptido-glycan is polymerized from a lipid phosphate-activated disaccharide by a protein that has a similar fold and mechanism to lysozyme is particularly intriguing.

Cellulose biosynthesis has also been heavily studied in plants, for which the current view mirrors that for bacteria: polymerization from UDP-glucose is attributed to the plant processive GT2 enzymes, which share signature motifs with their bacterial counterparts. The putative role of the cellulose KORRIGAN, a membrane-bound endo-1,4-β-glucanase, in cellulose synthesis is of interest, but a clear molecular function for this protein in cellulose biosynthesis in plants has yet to be found, although loss-of-function mutants have severe cellulose deficiency phenotypes.

Non-cellulose-synthesizing non-saprophytes. Of the bacterial taxa analysed, 121 (8% of the total examined) encode at least one cellulase but display no evidence of
a saprophytic lifestyle or the ability to synthesize cellulose (see Supplementary information S5 (table)). The existence of cellulase genes in these bacteria is extremely puzzling, as it is unclear what the function of these genes might be in these organisms. One example of such an organism is *M. tuberculosis*, which encodes two active cellulases and one active cellulose-binding protein regardless of the strain\(^3\). Two of these three proteins have secretion signal peptides. Likewise, all strains of the organism are involved in plant cell wall degradation — most recently, biomass conversion — but the role of intra-amoebal cellulases and one active cellulose degradation has not been conserved since the speciation events that gave rise to *M. tuberculosis* and the six other, closely related mycobacterial species that also cause tuberculosis in humans and animals (collectively, these seven mycobacteria are called the tuberculous mycobacteria). Furthermore, this protein and the two other proteins of *M. tuberculosis* that putatively target cellulose — encoded by Rv0062 (REF. 15) and Rv1987 (REF. 17) — displayed the expected activity on production in *E. coli*, and the other six tuberculous mycobacteria also encode these three cellulose-targeting proteins\(^16\). Together, these findings indicate that there has been evolutionary pressure to maintain the genes encoding these proteins.

*L. pneumophila* is an intracellular pathogen\(^36\) that lives in water-borne free-living amoebae\(^38\) and has evolved to resist macrophages after intra-amoebal ‘training’ (REFS 18, 38). *M. tuberculosis* is also an intracellular pathogen that can reside within host macrophages\(^39,40\), and it was recently shown to survive within amoebae following experimental infection\(^41\). Although the role of intra-amoebal survival in the life cycle of tuberculous mycobacteria remains purely hypothetical, it is possible that the cellulases in *M. tuberculosis* and *L. pneumophila* target the cellulose produced when amoebae encyst. Contrary to the non-tuberculous species *Mycobacterium avium*, the tuberculous mycobacteria *M. tuberculosis* and *Mycobacterium bovis* bypass the amoebal cyst through an unknown mechanism after internalization into the amoebal trophozoite\(^41\). The amoebal cyst wall is a simple structure consisting of an outer, proteinaceous ectocyst and an inner, cellulose-containing endocyst, and bacteria do not require a complex arsenal of enzymes to degrade this wall. Although an obvious hypothesis is that the cyst cellulose is broken down by the bacteria to generate glucose, more intricate roles for this breakdown cannot be excluded; for example, this degradation may enable the bacteria to detect amoebal cellulose synthesis at an early stage to trigger escape mechanisms before encystment.

Other bacteria besides *M. tuberculosis* and *L. pneumophila* can survive or multiply in amoebae, including *Vibrio cholerae*\(^42\), related *Vibrio* spp. and *Listeria monocytogenes*\(^43\). Together, these examples suggest a novel role for bacterial cellulases in the interactions of bacteria with the cellulose derived from amoebal encystment. The data discussed here could prompt additional studies on bacterium–amoeba relationships for bacterial species with one or more cellulases. However, for many bacteria, interactions with amoebae have not been documented in the published literature, and the role of cellulases in these organisms remains enigmatic (see Supplementary information S5 (table)).

**Conclusions**

We have tried to show that, contrary to popular acceptance, the presence of a cellulase-encoding gene in the genome of an organism does not automatically imply that the organism is involved in plant cell wall degradation. Until now, cellulases of saprophytic organisms have received attention mostly for their potential in industrial processes — most recently, biomass conversion — but the role of cellulases in non-saprophytic organisms is a novel

---

### Table 1 | The bacteria encoding the most cellulolytic enzymes in various taxa*

<table>
<thead>
<tr>
<th>Name (GenBank accession)</th>
<th>Number of cellulase genes</th>
<th>Number of hemicellulase genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phylum Actinobacteria</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptomyces bingchongensis str. BCW-1 (CP002047)</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>Amycolatopsis mediterranei str. U32 (CP002000)</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>Streptomyces scabiei str. 87.22 (FN554889)</td>
<td>20</td>
<td>33</td>
</tr>
<tr>
<td>Catamulspora acidiphila str. DSM 44928 (CP001700)</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>Actinosynnema mirum str. DSM 43827 (CP001630)</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td><strong>Phylum Gammaproteobacteria</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terehindibacter turnerae str. T7901 (CP001614)</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>Saccharophagus degradans str. 2-40 (CP000282)</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>Cellibrio japonicus str. Ueda 107 (CP000934)</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td><strong>Phylum Firmicutes (class Clostridia)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clostridium thermocellum str. ATCC 27405 (CP000568)</td>
<td>32</td>
<td>19</td>
</tr>
<tr>
<td>Clostridium thermocellum str. DSM 1313 (CP002416)</td>
<td>32</td>
<td>20</td>
</tr>
<tr>
<td>Clostridium cellobiolyticum str. 7413B (CP002160)</td>
<td>29</td>
<td>24</td>
</tr>
<tr>
<td>Ruminococcus albus str. 7 (CP002403)</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Ruminococcus sp. 18P13 (FP929052)</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td><strong>Phylum Bacteroides</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytophaga hutchinsonii str. ATCC 33406 (CP000383)</td>
<td>19</td>
<td>8</td>
</tr>
</tbody>
</table>

*Only bacteria with 15 or more cellulase genes are included.*
problem, the answer to which is emerging from our analyses of genomic data. These analyses show that many nonsaprophytic organisms encode a cellulase which, counterintuitively, plays a part in cellulose biosynthesis, revealing that cellulases are not restricted to cellulose degradation. The exact role of cellulase in this process remains unclear, despite the potential interest in controlling or engineering cellulose synthesis to produce materials with novel properties or enhanced biodigestibility. Furthermore, the cellulases that are present in human pathogens may reflect an intra-amoebal lifestyle and a requirement to degrade amoebal cellulose. In light of this genomic analysis, we suggest that cellulytic bacteria can currently be classified into four categories: those that use cellulases to feed on plant cell walls, those that use cellulases for cellulose biosynthesis, those that do both and those, like *L. pneumophila*, that may at one stage in their life cycle reside inside amoebae (FIG. 3a). Given the important societal and health issues related to the persistence of both mycobacteria and *Legionella* spp., the observation that cellulases may be implicated in the mechanisms of persistence for these organisms is thought-provoking and will perhaps lead to the generation of novel strategies to combat these pathogenic organisms.

Figure 4 | Polysaccharide biosynthesis operons and cellulase fusion proteins. a | Polysaccharide biosynthesis operons in bacterial genomes, including the bacterial cellulose synthesis (bcs) operon of *Escherichia coli* str. K12 and analogous operons in other species. The glycoside hydrolase (GH)-encoding gene of the bcs operon, bcsZ, is from the GH8 family; other GH families are represented in analogous operons. Similarly, each operon contains a processive glycosyltransferase 2 (GT2) family-encoding gene, which in the bcs operon is bcsA. GenBank, Gene or Protein database accession are given in brackets for each GH family gene (or the encoded protein). | Polysaccharide biosynthesis operons and cellulase fusion proteins.

ANALYSIS


Acknowledgements

F.M.M. was funded by La Fondation Infectiopole Sud, France. G.J.D. is a Royal Society Wolfson Research Merit Award recipient.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

CAzY: http://www.cazy.org/

enzymes: http://www.enzyme.org/


SUPPLEMENTARY INFORMATION

See online article: S1 (table) | S2 (table) | S3 (table) | S4 (table) | S5 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF