Inter- and intra-domain horizontal gene transfer, gain–loss asymmetry and positive selection mark the evolutionary history of the CBM14 family

Ti-Cheng Chang and Ioannis Stergiopoulos

Department of Plant Pathology, University of California Davis, CA, USA

Keywords
birth-and-death evolution; carbohydrate-binding modules; chitin; diversification; ligand specificity

Correspondence
I. Stergiopoulos, University of California Davis, Department of Plant Pathology, One Shield Avenue, Davis, CA 95616-8751, USA
Fax: +1 530 752 5674
Tel: +1 530 400 9802
E-mail: istergiopoulos@ucdavis.edu

(Received 25 November 2014, revised 1 March 2015, accepted 3 March 2015)
doi:10.1111/febs.13256

Protein–carbohydrate interactions are ubiquitous in nature and at the core of many physiological processes of profound importance to health and disease. Specificity in protein–carbohydrate interactions is conferred by carbohydrate-binding modules (CBMs) that can accurately discriminate among the multitude of saccharides found in nature, thus targeting proteins to their particular substrates. Family 14 carbohydrate-binding modules (CBM14s), more specifically, are short modules that bind explicitly to chitin, the second most abundant carbohydrate in nature. Although considerable effort has been placed in elucidating protein–carbohydrate interactions at the molecular level for biological and biotechnological applications, in contrast the evolutionary relationships among these modules are minimally understood. Using the CBM14 family as an example, here we describe one of the first global molecular evolutionary analyses of a CBM family across all domains of life, with an emphasis on its origin, taxonomic distribution and pattern of diversification as a result of gene and module duplication, and positive selection. Our genome-wide searches recovered an impressive number of CBM14s from diverse lineages across nearly all domains of life. However, their highly disseminated distribution in taxa outside the Opisthokonta group strongly suggests a later evolutionary origin and elevated rates of inter- and intra-domain horizontal gene transfer. Moreover, accelerated rates of asymmetric gains and losses reveal a dynamic mode of birth-and-death evolution, whereas positive selection acting on paralogous CBM14-containing proteins suggest changes in substrate specificity and an increase in the functional promiscuity of this ancient CBM family. The importance of these results is discussed.

Introduction

Carbohydrate-binding modules (CBMs) are ubiquitous molecules in nature that bind to polysaccharides or oligosaccharides of varying chain lengths [1,2]. They are frequently present as discrete non-catalytic components of modular carbohydrate-active enzymes (CAZy) that promote the interaction of the enzyme with the target polysaccharide substrate, thus increasing the efficiency of catalysis [1,3]. Lectins are carbohydrate-binding proteins that lack catalytic activity but often contain one or several CBMs in tandem repeats [4,5]. At least 71 families of CBMs are currently described in the CAZy database that display highly diverse ligand specificities, thereby enabling these modules to be at the very heart of critical biological

Abbreviations
CAZy, carbohydrate-active enzymes; CBM, carbohydrate-binding module; CBM14, family 14 carbohydrate-binding modules; HGT, horizontal gene transfer; LOCA, last opisthokont common ancestor; LRT, likelihood ratio test; ML, maximum likelihood; PAD, peritrophin-A domain.
processes, such as viral and microbial infection, innate immunity, tumor spread and others [2,3,6]. Family 14 carbohydrate-binding modules (CBM14s) (PFam ID, PF01607; SMART, SM00494), in particular, are short modules of approximately 70 residues that are found in a diverse group of organisms, including eukaryotes, viruses and bacteria [7,8]. To date, chitin and soluble chito-oligomers are the only reported ligand targets of CBM14s, and thus they are frequently connected to catalytic domains with chitinolytic activity [8–10]. While the presence of CBM14s in modular chitinases plays a key role in enzyme targeting and function, several members of the CBM14 family are also frequently encountered as chitin-binding lectins [7,8,11]. At present, more than 1500 polypeptide sequences of CBM14s have been deposited in the CAZy database, with the vast majority originating from eukaryotes (1252 sequences) and only a limited number from other domains of life, such as viruses (246 sequences) and bacteria (one sequence). However, despite having over 1500 identified members, family 14 remains one of the least characterized CBM families. In insects, the CBM14 is also widely known as the peritrophin-A domain (PAD), as it is frequently found in chitin-binding proteins of the peritrophic matrix [12–14]. Examination of PAD proteins from insects showed that their numbers and domain architectures can be very diverse [14,15]. Most PAD proteins are also predicted to have cleavable signal peptides, suggesting that they are secreted proteins that interact with chitin in extracellular matrices. An evolutionary analysis of insectal peritrophic matrix proteins with PAD domains (PAD-PMPs) revealed an equally distant relationship between chitinases and PAD-PMPs without separated clusters, suggesting a potential shared ancestry of a protein with a single PAD. In addition, phylogenetic analysis showed clustering of PAD-PMPs with multiple repeats, suggesting that they may have arisen from recent gene duplication and crossover events [16]. However, apart from these studies referring to specific insect lineages, the global evolution of the entire CBM14 family remains elusive.

Structural information about the CBM14 is also currently restricted only to tachycitin, a small antimicrobial lectin-like protein from the horseshoe crab Tachypleus tridentatus [7]. The elucidation of the solution structure of tachycitin by NMR revealed that the putative chitin-binding site of this lectin consists primarily of six cysteine residues that form three disulfide bridges and several aromatic residues that most probably play a pivotal role in substrate-binding and specificity [8]. Tachycitin also exhibits a similar disulfide bond pattern and local sequence similarity in the putative chitin-binding domain with another CBM14 family member, Avr4 from the fungal pathogen Cladosporium fulvum [11]. Avr4 is a small 135 residue secreted protein that binds specifically to chitin present in fungal cell walls, thus protecting them against hydrolysis by plant chitinases during infection of the tomato host [17]. The tertiary homology presumably shared by tachycitin and Avr4 suggests that members of the CBM14 family may adopt a common structural fold [11]. To date, functional orthologs of Avr4 have been identified in a few fungal species of Dothideomycete and, despite their low overall homology (30–40%), they all bind chitin and share a similar function in protecting fungal cell walls against plant chitinases [18].

Although a significant amount of effort is currently being placed on deciphering the biochemical bases of CBM-ligand binding and recognition, in stark contrast the evolutionary relationships among these modules, both between as well as within the different CBM families, are poorly understood [1]. Also, most phylogenetic and molecular evolutionary studies with CBMs have been restricted within specific lineages and species, merely aiming to address questions specific to the organismal biology from which the respective molecules are found [6,19,20]. In this study, we provide an evolutionary framework of the entire CBM14 family across all domains of life in order to obtain a global understanding of the evolutionary dynamics of a CBM family. More specifically, we first use a combination of exhaustive systematic and iterative searches to uncover the true breadth of CBM14 family members present across all domains of life and to define signature motifs within the family. We then examine the phylogenetic distribution of the family members, infer ancestral copy number, and test hypotheses regarding its origin and pattern of evolution across different lineages and taxa. We finally investigate instances of diversification facilitated by adaptive evolution and a variety of molecular mechanisms. Overall, the analysis elucidates how the CBM14 family arose, evolved through horizontal gene transfer (HGT) and multiple expansions and contractions in a lineage-specific manner, and potentially diversified through positive selection. In addition to highlighting the evolutionary history of an important CBM family, the data presented here provide a framework for understanding the complex evolutionary history of other CBM families as well.
Results and Discussion

Homology levels among retrieved CBM14 sequences reveal high rates of sequence diversification and a few conserved motifs

A total of 3432 proteins that comprise 7660 homologous CBM14 sequences were identified based on our searches (Doc. S1). This number is significantly larger than the ~1500 entries of CBM14 sequences currently present in the CAZy database and slightly larger than the ~6920 entries deposited in the Pfam database (Pfam ID: PF01607). These differences are mainly a consequence of the different number of genomes and/or timeframe of curation processes in different databases.

CBM14 sequences are highly diverged, with an average amino acid sequence similarity of 16%. Despite the low similarity, hierarchical clustering based on pairwise e-values computed after random shuffling of the CBM14 sequences and visualized on a heatmap displayed a central ‘hot-block’ with e-values ≤ 1E-5 (Fig. S1). Nearly all the identified CBM14s shared a significant reciprocal similarity to the sequences within this block (e-value 9E-7), thereby supporting the homology of the identified CBM14 sequences. Furthermore, although less strongly demarcated, five blocks of high correlation (blocks 1–5) that could represent subfamily divisions were identified within the hierarchical clustering tree, four of which were located within the central hot-block. Blocks 1 and 2 form the two largest groups and are composed of 2009 and 988 CBM14 sequences, with 25.5% and 23.2% intra-block protein sequence similarity, respectively. Motif analysis in these two groups identified several conserved residues that were distributed between the six conserved cysteine (C) residues that denote the CBM14 [14] and which were arranged in short conserved sequence patterns or motifs (Fig. S2). For example, in groups 1 and 2 conserved residues between the second and third cysteine formed the consensus motif C-X-K-[F/Y]-Y-X-C, between the third and fourth cysteine motif C-X2-G-X6-C, and between the fourth and fifth cysteine motif C-P-X-G-L-X-F-[N/D]-X5-C. These motifs are well conserved in groups 1 and 2, and appear as the main core motifs in the entire CBM14 family. The remaining three groups contained fewer CBM14 sequences, with 496 (24.3% similarity), 429 (24.8% similarity) and 426 (29.5% similarity) members, respectively. Motifs present in groups 3 and 4 generally shared a similar pattern to the first two groups but with a few discrepancies. For example, an additional conserved alanine (A) residue was present between the fourth and fifth cysteine in group 3, leading to motif C-P-X-G-L-X-F-A-[N/D]-X5 (Fig. S2). The overall spacing distance between the six conserved cysteine residues is also relatively conserved, with one or two amino acid space differences between four of the five groups. The motif of group 5 exhibited a relatively different pattern from the four described groups, mainly due to its diverged cysteine spacing pattern and sequence composition. The similarity between group 5 and the other groups ranged from 19% to 22% but nevertheless the described core motifs of C-X2-G-X6 and C-P-X-G-L-X-F-[N/D]-X5-C are still present in group 5. The other minor groups in the clustering tree shared similar motif patterns and displayed ~20–30% intragroup and ~5–20% intergroup similarity (Figs S1 and S2). Overall, these results indicate that, despite the high sequence diversification observed among members of the CBM14 family, a small set of residues seem to have been conserved during evolution and thus might play essential roles in ligand-binding and recognition and/or protein conformation and stability.

The broader taxonomic distribution of CBM14 family members across the entire tree of life suggests the presence of multiple HGT events in its evolutionary history

Our iterative searches enabled the identification of CBM14s from 382 species that were distributed across most major domains in the tree of life (Fig. 1, Doc. S1). The majority of species with CBM14s was present within the eukaryotic clade (315/382, 82.5%) and spread across a range of phylogenetically diverse lineages. From the six major supergroups that compose the eukaryotic domain of life [21], namely Opisthokonta, Amoebozoa, Archaeplastida (Plantae), Chromalveolata, Rhizaria and Excavata, CBM14s were predominantly present within the metazoan (256 species; 7388 modules) and fungal kingdoms (51 species; 86 modules) of Opisthokonta, and rarely in Amoebozoa (two species; six modules), Archaeplastida (three species; eight modules) and Chromalveolata (three species; 31 modules). Also, CBM14s were totally absent in available genome sequences from Rhizaria and Excavata (Fig. 1). In non-eukaryotes, CBM14 family members were absent in archaean and present in only six species within the bacterial domain that present diverse bacteria phyla, including the gram-negative Proteobacteria (sections of Betaproteobacteria and Gammaproteobacteria (two species; seven modules) and Bacteroidetes (one species; two modules)) as well as the gram-positive...
Actinobacteria (two species; three modules) (Fig. 1). Although bacteria do not produce chitin, it is possible that chitin-binding proteins have been acquired in order to enhance the function of bacterial chitinases for defense against chitinous pathogens, and/or promote the biodegradation of insoluble chitin for use as a nutrient source [22]. 124 members of the CBM14 family were also recovered from diverse double-stranded DNA viruses (61 species), including viruses from three of the four genera in the family of Baculoviridae [i.e. Alphabaculovirus (38 species), Betabaculovirus (10 species) and Gammabaculovirus (three species)]. The presence of chitin-binding domains in this insect-specific pathogen family strongly suggests a role in insect infectivity, such as the attachment of the occlusion-derived virions to the chitin-rich peritrophic matrix or the enhancement of the chitinolytic activity of baculoviral secreted chitinases [23]. CBM14s were also recovered from several strains of the Paramecium bursaria chlorella virus (chloroviruses; family Phycodnaviridae) that infects chlorella-like green algae [24]. Notably, chloroviruses are the only non-eukaryotic organisms known to synthesize chitin, in which it plays an important role in the infection cycle [25].

The widespread presence of CBM14 family members mainly within the metazoan and, to some extent, the fungal kingdoms coincides with the widespread distribution of chitin in metazoans and fungi. This suggests that the CBM14 possibly emerged around the divergence time of the Opisthokonta group within the Eukaryotic domain, before the split of the fungal and metazoan lineages. Notably, the appearance of the CBM14 also coincides closely with the emergence of chitin synthases (Chs) in nature, which are proposed to have emerged once the plant kingdom diverged within the Eukaryotic domain [26]. Additionally, the rare presence and punctate distribution of CBM14 family members in diverse phyla and species outside the Opisthokonta group suggests that CBM14s are likely to have been introduced in these lineages independently by multiple HGT events. Despite the fact that inter-domain HGT is generally considered as a...
relatively rare event [27,28], the alternative scenario of vertical acquisition from a common ancestor would seem less parsimonious, considering the substantial number of gene loss events that would need to take place in order to explain the aberrant distribution of CBM14-encoding genes in the array of taxonomically divergent species outside the Opisthokonta group.

To test this hypothesis, we performed a phylogenetic analysis of the entire CBM14 family using a maximum likelihood (ML) approach and compared it to the corresponding reference species tree obtained from the Tree Of Life (sTOL) [29] and NCBI taxonomy databases (Fig. S3). It should be noted that since the CBM14 family has undergone substantial sequence diversification in its evolutionary history, the phylogenetic tree produced had limited bootstrap support, especially within the deep nodes of the phylogeny. This indicates that the historical associations of the CBM14 family members cannot be resolved with confidence beyond the level of closely related species (Fig. S3). Nevertheless, despite the weak phylogenetic signal, comparison of the CBM14 family member phylogeny to the species phylogeny for relatively well supported clades identified several cases of incongruent topologies, as CBM14 family members were highly scrambled with respect to the expected organismal phylogeny. CBM14s from prokaryotes, for instance, were highly scattered in distinct regions of the phylogenetic tree and were frequently nested within clades that otherwise consisted of CBM14s from eukaryotes (Fig. S3). Even though such topological discordance between the CBM14 family member tree and the species tree could be due to ancient duplication events and the coexistence of multiple CBM14 copies in the commonote (the common ancestor of eukaryotes and prokaryotes), such a scenario would imply numerous independent losses of paralogous gene copies in multiple eukaryotic and prokaryotic lineages, thus making an origin via vertical descent very unlikely [28,30].

To further test the hypothesis of eukaryote-to-bacteria HGT, we used conSEL and PRUNIER to examine whether for clades with strong bootstrap support the current ML topologies are significantly different from constrained topologies, in which the eukaryotic proteins from the clades that nested bacterial proteins form monophyletic groups including or not the bacterial proteins [31–33]. The tests showed that in at least one such case the likelihood of the constrained topologies was significantly different from the ML topology, thus rejecting at the 5% significance level the alternative non-HGT topology. More specifically, the CBM14 present in the bacterial species Sphingobacterium spiritivorum (UniProt ID: C2G1B4) was clustered with a CBM14 from the blacklegged tick Ixodes scapularis (UniProt ID: B7P6A4) and both were further nested within a clade that consisted of CBM14s mainly from Arthropoda (Fig. 2, Table S1). Although this relationship was moderately supported (bootstrap support of 69) compared to an alternative topology in which the insect taxa were forced to form a monophyletic group with the exclusion of the module from S. spiritivorum, the currently observed topology was significantly supported at the 5% significance level by all the topology tests in conSEL. The Shimodaira–Hasegawa test, sh (the Shimodaira–Hasegawa test) and wkh (the weighted Shimodaira–Hasegawa test) P values 0.974, au (the approximately unbiased test) P value 0.9802 (Fig. 2).

This strongly suggests that S. spiritivorum could have been the recipient of a horizontally transmitted CBM14 from donor insect taxa. In all other cases of suspected eukaryotes-to-bacteria HGT examined, the tree topology tests failed to reject the null hypothesis of a vertical origin. Such lack of statistical support for HGT events, although suggestive of vertical inheritance, could be caused by the short sequence lengths and fast rate of evolution of the CBM14 family that have rapidly degenerated CBM14 sequences and produced phylogenies with limited support [33].

Similar putative cases of HGT were also identified when considering inter-domain eukaryote-to-eukaryote transfers. Within the Archeplastida, for example, CBM14s were identified in the three relatively distant green algae species Chlamydomonas reinhardtii, Chlorella variabilis and Volvox carteri f. nagariensis but were absent from higher plants, suggesting that the CBM14 was introduced in green algae by HGT. This would be in agreement with previous reports that lateral gene transfers must have taken place between green algae and animals [34,35]. Moreover, the CBM14s from the three green algae species exhibited a punctate taxonomic distribution within other eukaryotic lineages that suggests HGT. The CBM14 from, for instance, C. variabilis (UniProt ID: E1ZLJ9) was clustered with relatively strong support (bootstrap of 94) with CBM14s from the nematode species Caenorhabditis reman (NCBI ID: XP_003115066) and Caenorhabditis briggsae (NCBI ID: XP_002640431) (Fig. S4, Table S1). All tree topology tests implemented in conSEL favored at the 5% significant level (kh, sh and wkh P values 0.946, au P value 0.962) the placement of the C. variabilis CBM14 within the Caenorhabditis group and rejected the alternative constrained topology of this module branching out from the nematode group, thus supporting the hypothesis of a putative HGT event. Similar observations and postulates about
HGT could also be made for the CBM14s present in the amoebozoan species *Dictyostelium purpureum* (NCBI ID: XP_003284035) and *Polysphondylium pallidum* (NCBI ID: EFA74810), as well as the diatom species *Thalassiosira pseudonana* (UniProt ID: B8C632) for which the branching topologies that nested the CBM14 sequences from these species within different eukaryotic clusters consistently received stronger statistical support compared to the constrained alternative topologies (Fig. S4).

The above cases are examples of suspected eukaryote-to-prokaryote and inter-domain eukaryote-to-eukaryote HGT in the CBM14 family that, if real, would imply a surprisingly high rate of lateral transfers in the evolutionary history of this family. However, these results should be taken with caution as the phylogenetic analysis identified several more cases of topological incongruences between the CBM14 family member tree and the species tree that extended within Opisthokonta lineages as well. Although most of these cases could represent phylogenetic artifacts caused by the extreme heterogeneity of the amino acid composition displayed by CBM14 sequences and/or compositional bias, alternatives to HGT scenarios, such as ancient paralogy coupled with unappreciated high rates of gene duplication and loss or independent/convergent evolution in different lineages, cannot yet be unambiguously excluded and could possibly account for some of the unexpected phylogenetic relationships within the CBM14 family. It is worth mentioning here that, despite their high sequence diversity, the topology of CBM14 family members from different fungal species formed a monophyletic group that did not include sequences from other domains of life. Findings like these imply a single origin for the fungal CBM14s and are in stark contrast to the punctate phylogenetic positioning of CBM14s from non-Opisthokonta species within the Opisthokonta group.

**Frequent and abundant lineage-specific gains and losses shape the evolutionary dynamics of the CBM14 family within the metazoan and fungal lineages**

Since its initial appearance within the Opisthokonta group, the CBM14 family has experienced several lineage-specific expansions and contractions in its evolutionary history, thus revealing a complex and dynamic mode of molecular evolution that is consistent with the birth-and-death model of gene family evolution.
[36,37]. More specifically, ancestral state reconstruction suggested that possibly between three and four CBM14s could have been present in the last opisthokont common ancestor (LOCA), whereas the subsequent evolution of the CBM14 family consisted of multiple episodes of differential expansion, contraction and even extinction in several lineages (Fig. S5). Overall, a high number of changes in terms of gains and losses in CBM14s were observed within internal (228 changes) and terminal (305 changes) branches of the Opisthokonta phylogeny, indicating that both lineage- and species-specific adaptations shape the evolutionary trajectory of the CBM14 family. Also, in both internal and terminal branches of the phylogeny, an equally high number of gains (116 and 138, respectively) and losses (112 and 167, respectively) were observed, suggesting that they are frequent events in CBM14 family evolution that occur with almost equal frequency. Despite, however, the overall similarity between internal and terminal branches of the phylogeny, the gain/loss rate in CBM14s varied substantially among the phylogenetic lineages and was mostly asymmetric, as sister lineages and extant taxa often exhibited considerable differences in the net number of CBM14s encoded in their genomes (Fig. S5). Taken together, these observations suggest lineage-specific adaptations and discrete evolutionary trajectories of the CBM14 family in different taxa.

Within Holozoa, for example, one of the two major subgroups that comprise the Opisthokonta supergroup, CBM14s were abundantly present in Metazoans but were absent in Choanoflagellates (the sister clade to the Metazoans) as well as in the early diverging Holozoan lineages of Filasterea and Ichthyosporea [38]. The expansion of the CBM14 family in metazoans was not uniform across all lineages but mainly took place in Ecdysozoa, and particularly in Nematoda (roundworms) and the insectal clades of Drosophila (fly), Culicidae (mosquitoes) and Apocrita (wasps, bees and ants) (Fig. S5). The median number of CBM14s present in each of these four lineages was 32, 217, 66.5 and 46, respectively, which is considerably higher than almost all other lineages within Opisthokonta and the 7.9 median number of CBM14s averaged across all other major lineages in the tree of life (Fig. S6). Notably, the large expansion of the CBM14 family in these lineages took place on both the internal and the terminal branches of the phylogeny, indicating a continuous increase in family size rather than isolated species-specific bursts of gains in extant species. At the same time, however, in several taxa there was also a loss of ancestral CBM14s that were gained before the point of duplications, leading to the contraction, or even complete extinction, of the family in some species. A notable example of this asymmetric pattern of gains and losses in CBM14s is the consecutive gain of these modules at internal branches of the Dipteral phylogeny that lead to the Drosophila clade and the further episodic expansion of the family in some of the Drosophila species (Fig. S5). For instance, in D. melanogaster, a total of 475 CBM14s were identified, more than double the amount from the median number of 217 CBM14s present among the 15 Drosophila species that were included in this study. Similar patterns were also observed in the Arthropoda lineages of Culicidae (mosquitoes) and Apocrita (wasps, bees and ants) as well as in Nematoda, the sister clade to the Arthropods (Fig. S5).

In contrast to Nematodes, Drosophila, Culicidae and Apocrita, most other Ecdysozoa and Protostome in general lineages exhibited less gain or loss in CBM14s, with only occasional large episodic expansions in some of the extant species (Fig. S5). This is also the case in Lophotrochozoans, the third largest group in Protostomia next to the Arthropods and Nematodes, in which between one and four CBM14s were observed per species, with sporadic large episodic species-specific expansions. Notably, a recent study indicated that Lophotrochozoans exhibit the largest pool of Chs genes within the metazoan group [39]. A similar tempo of evolution was also present within the metazoan lineages of Deuterostomia, where mostly small net changes in the number of CBM14s present in extant taxa were observed. The only notable exception is the large lineage-specific expansion of the CBM14 family in some of the lower chordates, including the three species of Tunicata included in this study, which exhibited a median number of 41 modules, and the Cephalochordate lancelet Branchiostoma floridae, in which a total of 108 CBM14s were observed, the highest number of CBM14s observed so far in the species of Deuterostomia (Fig. S5). The occasional expansion of the CBM14 family was also observed in some vertebrate species of Chordata animals, including the opossum Monodelphis domestica, the threespine stickleback fish Gasterosteus aculeatus as well as humans, where 16, 16 and 15 CBM14s were found, respectively. The presence of chitin-binding proteins in Deuterostomes would perhaps corroborate the recent discovery of Chs genes in this lineage, which differs from previous beliefs that the ability to produce chitin was lost at the root of the Deuterostome lineage and suggests that chitin could be more widely distributed in Deuterostomia than initially thought [39].

Finally, in contrast to Holozoa, the fungal lineage is the most interesting lineage from the perspective of
putative loss, as the CBM14 family was entirely lost from most lineages within the fungal kingdom, leaving only a small number of 71 surviving members distributed among 51 of the 120 fungal species (97 Ascomycetes, 18 Basidiomycetes, three Zygomycetes, two from other early divergent fungal lineages) that were included in this study (Fig. S5). Almost all species with CBM14s belong to the phylum of Ascomycota (43 species), with the exception of one species of Basidiomycota and four species of Zygomycotina that also possess CBM14s. The high number of CBM14s within the Ascomycota would coincide with the important structural role of chitin in the cell walls of higher fungi, in contrast to Zygomycetes, for example, which synthesize mostly chitosan, a de-acetylated form of chitin [40–42]. In Basidiomycetes, however, the CBM14 family has been almost entirely lost, despite the fact that they utilize chitin as one of the main structural cell wall polysaccharides [42]. Within Ascomycota, CBM14s were only found in the species of the subphylum Pezizomycotina and were most abundant within the Aspergillaceae family of Eurotiomycetes (24 of 35 species) as well as within the Mucosphaerellaceae family of Dothideomycetes (14 of 28 species). Outside these two classes, CBM14s were recovered from only a handful of species in other major lineages within Pezizomycotina, including Leotiomycetes (three of 11 species) and Pezizomycetes (two of seven species). The high loss rate of the CBM14 family in fungi was also evident when considering the number of modules present in each genome. At the base of the fungal lineage only a single CBM14 is predicted to be present, which represents a loss of two to three CBM14s compared to the LOCA. The median number of CBM14s present in fungi is 1.7, as most fungal species had one (29 species) or two modules (19 species) present in their genomes, and only one species was identified with three CBM14s and one species with four CBM14s. Finally, one case of a large episodic expansion of the CBM14 family was observed in the saprobic Pezizomycete Pyronema omphalodes (syn. P. confluens), since 13 CBM14s were found in the genome of this fungus (Fig. S5).

A variety of molecular mechanisms contribute to the expansion of the CBM14 family within Opisthokonta lineages

A closer examination indicated that the observed expansion of the CBM14 family in specific lineages and extant taxa was due to duplication of genes encoding CBM14-containing proteins as well as to within-protein tandem duplications of the CBM14. When considering differences in counts of CBM14-containing proteins among the different species, of the 382 species in all domains in the tree of life that possessed CBM14s, 39% (149/382) had only a single CBM14-containing protein while the remaining 61% (234/382) had two or more CBM14-containing proteins. Also, an extremely high variation in counts of CBM14-containing proteins among the different species was observed, with numbers varying from 1 to 177 proteins per species (median of three proteins per species) (Fig. S7). However, despite this notable variation from species to species, a positive correlation (Pearson’s correlation coefficient $R = 0.977$) existed between expansion of the CBM14 family and increases in the number of CBM14-containing proteins, most likely as a result of gene duplications. This was true, for instance, for the main metazoan clades in which a large expansion of the CBM14 family took place, including the clades of Nematoda ($R = 0.898$), Drosophila ($R = 0.997$), Culicidae ($R = 0.988$) and Apocrita ($R = 0.990$) within Protostomia, and the Tunicata ($R = 0.988$) clade within Deuterostomia (Fig. S8).

Next to an increase in the number of genes encoding CBM14-containing proteins, the expansion of the CBM14 family was positively correlated with an increase in the number of CBM14s per protein, suggesting that, in addition to gene duplication, the emergence of new protein architectures through CBM14 duplication and accretion has played a substantial role in the evolution of the CBM14 family. In total, 58% (1991/3432) of all the CBM14-containing proteins were found to have a single CBM14 within their sequences, while the remaining 42% (1441/3432) of the proteins contained from two to as many as 35 CBM14s. Stated differently, of the 7660 CBM14s that were identified, 5668 (73%) were found to be from proteins that contained at least two CBM14s within their polypeptide sequence, while the other 1992 domains (27%) were present in proteins with only a single CBM14. Most proteins with multiple CBM14s had two (469, 13.7%), three (492, 14.3%) or four (169, 4.9%) CBM14s but several proteins (311, 9.1%) were also identified with five or more CBM14s (Fig. S9).

The expansion of the CBM14 family through gene duplications and within-protein CBM14 duplications was evident in the Drosophila clade, which exhibited the largest expansion in the CBM14 family (Fig. 3). Changes in gene family size among Drosophila species were mainly due to an expansion in proteins with a single CBM14 in their polypeptide sequence, which are extremely abundant in D. melanogaster (91) and less in the other Drosophila lineages (Fig. 3).
However, given the existing variations in levels of genome completion and annotation among the different Drosophila species and beyond, caution needs to be drawn to the fact that the absence of proteins with a single CBM14 in some species could perhaps be ascribed to the incompleteness of the current genome assemblies and annotations. High variation in the number of proteins with multiple CBM14s and in the number of CBM14s per protein was also observed. In this respect, the median number of proteins with a single CBM14 in their structure among all 14 species of Drosophila included in this study was 34, which is much higher than the median number (two) of proteins with two or more CBM14s. In these proteins, CBM14 counts per protein varied from 1 to 21, indicating that, next to gene duplication, putative within-protein tandem duplication of CBM14s has played a significant role in the expansion of the CBM14 family in Drosophila species. Overall, although the completeness of the genome assembly and annotation may impact the observed variations, significant variations of the expansion patterns were still present among the species with well annotated genomes. Similar observations could also be made for almost all other eukaryotic lineages in which the CBM14 family has been expanded.

Positive Darwinian selection among paralogous CBM14-containing proteins further suggests functional diversification in the CBM14 family

The complex evolutionary history exhibited by the CBM14 family as a result of frequent gene duplications and within-protein tandem duplications of CBM14s suggests that members of this family may be subject to positive Darwinian selection and perhaps functional diversification as well. We investigated this possibility by searching for evidence of adaptive molecular evolution in a set of paralogous proteins with a single CBM14 in their sequence from the common fruit fly *D. melanogaster* and a similar set of proteins from the Ascomycete *P. omphalodes*. We selected these two species because they represent the metazoan and fungal lineages, respectively, which are the two main lineages in Opisthokonta where the CBM14 family has followed a slightly different evolutionary route. Second, these species were selected because both exhibited the largest expansion in the CBM14 family in their respective lineages, mainly as a result of abundant duplication and retention of genes encoding extracellular proteins with a single CBM14.

As selection is most frequently directed towards a few individual sites rather than the entire domain
sequences, we first used the ML codon-based site models implemented in CODEML [43,44] to determine instances of departures from the neutral model of molecular evolution and to identify positively selected sites. In D. melanogaster, the likelihood ratio test (LRT) comparing models M3 (discrete) and M0 (single rate) indicated a significant deviation from the neutral model of molecular evolution ($2\Delta L = 493.549$, $P < 0.0001$, df = 4), thus suggesting heterogeneous selection pressure among codon sites that could have promoted divergence of the CBM14 family members. To further identify specific codons evolving with $\omega > 1$, we tested whether models M2a (selection) and M8 (beta&$\omega$), which allow for positively selected sites, fitted the data better than their nested null models M1 (neutral) and M7 (beta), respectively [45]. LRTs showed that the positive selection model M2a was significantly better than the neutral model M1a at 1% cutoff ($2\Delta L = 17.532$, $P = 0.0002$, df = 2), thus revealing the presence of positively selected sites in the codon alignment (Table 1). Model M2a also suggested that 15.1% of the sites are evolving under positive selection with $\omega_2 = 74.37$, while the Bayes Empirical Bayes (BEB) estimation method [46] detected a total of seven positively selected amino acid sites, six of which (T2, E20, S21, H24, S25 and H38, denoted based on UniProt ID: A1ZA23) received strong statistical support with a posterior probability (pp) > 0.99 (Table 1). Estimates under model M8 also suggested that a small proportion of just 3.1% of the sites are under positive selection with $\omega_2 = 13.6$, although based on the LRT model M8, was not able to reject M7, indicating that it is unlikely that these sites are under strong selection. Similar results were obtained when examining instances of adaptive evolution among paralogous CBM14-containing proteins in P. omphalodes, in which case the site model tests based on LRTs were also not able to provide strong evidence for positive selection (Table S2).

The lack in most cases of statistical support for positive selection would hint towards purifying selection acting on some of the paralogs. However, it is unlikely that positive selection would have operated on a large number of sites over prolonged periods of time. Rather, adaptive evolution or relaxed selection in some of the paralogs might have been in effect in the early stages following gene duplication and/or during a certain evolutionary window, whereas purifying selection may have acted later on to stabilize the new function [47]. To investigate this possibility, we examined the selection pressure by traversing all the internal and terminal branches of the CBM14 tree in both data sets using the branch-site models implemented in PAML [48]. In D. melanogaster, a total of 121 branches were tested for positive selection by the LRT conducted between model A (positive selection) and A-null (relaxation). The LRT results indicated that the $2\Delta L$ of 15 branches, including eight genes and seven internal branches, were all greater than the suggested threshold of 3.84 ($P < 0.05$) and thus subject to accelerated paces of evolution (Fig. S10, Table S3) triggered by positive Darwinian selection, relaxation of selection pressure, or a combination of both. Even after correcting for multiple testing, three branches remained

### Table 1. Log-likelihood values and parameter estimates under the site models implemented in CODEML for the selected CBM14 sequences from Drosophila melanogaster. $p$ refers to the number of parameters in the $\omega$ distribution. dN/dS is the average dN/dS value over codons. The positive selected sites column indicates the amino acid sites found to be under positive selection with posterior probabilities $> 95\%$ (bold) in the BEB analysis. Sites are annotated based on the CBM14 protein UniProt ID: A1ZA23_DROME from D. melanogaster. $2\Delta L$ is the likelihood ratio tests comparing a null and positive selection models (M1a versus M2a, M7 versus M8). df is the degree of freedom in the site models of CODEML.

<table>
<thead>
<tr>
<th>Model</th>
<th>$p$</th>
<th>$\ln L$</th>
<th>dN/dS</th>
<th>Parameters</th>
<th>Positively selected sites</th>
<th>$2\Delta L$</th>
<th>df</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>One ratio (M0)</td>
<td>1</td>
<td>-4552.703</td>
<td>0.119</td>
<td>$\omega = 0.119$</td>
<td>None</td>
<td>493.549</td>
<td>4</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Discrete (M3)</td>
<td>5</td>
<td>-4305.928</td>
<td>0.140</td>
<td>$P_0 = 0.143$, $\omega_0 = 0.003$, $P_1 = 0.343$, $\omega_1 = 0.068$, $P_2 = 0.514$, $\omega_2 = 0.226$</td>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nearly neutral (M1a)</td>
<td>2</td>
<td>-4426.122</td>
<td>0.705</td>
<td>$P_0 = 0.324$, $\omega_0 = 0.091$, $P_1 = 0.676$, $\omega_1 = 1.000$</td>
<td>Not allowed</td>
<td>17.532</td>
<td>2</td>
<td>0.0002</td>
</tr>
<tr>
<td>Positive selection</td>
<td>4</td>
<td>-4417.355</td>
<td>11.91</td>
<td>$P_0 = 0.144$, $\omega_0 = 0.018$, $P_1 = 0.706$, $\omega_1 = 1.000$, $P_2 = 0.151$, $\omega_2 = 74.37$</td>
<td>T2 (1.000), S3 (0.859), E20 (0.998), S21 (1.000), H24 (0.974), S25 (0.996), H38 (0.999)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta (M7)</td>
<td>2</td>
<td>-4305.594</td>
<td>0.126</td>
<td>$P = 0.542$, $\omega = 3.601$, $P_0 = 0.969$, $\omega_0 = 0.567$, $q = 4.368$, $P_1 = 0.031$, $\omega = 13.589$</td>
<td>Not allowed</td>
<td>0.806</td>
<td>2</td>
<td>0.668</td>
</tr>
<tr>
<td>Beta&amp;$\omega$ (M8)</td>
<td>4</td>
<td>-4305.191</td>
<td>0.524</td>
<td>$P_0 = 0.542$, $q = 3.601$, $P_0 = 0.969$, $\omega_0 = 0.567$, $q = 4.368$, $P_1 = 0.031$, $\omega = 13.589$</td>
<td>S21 (0.534)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
significant at the 5% significance level (Table S3). From the 15 branches subject to possible positive selection, six different residues (Y6, D8, E13, H16, W26 and P29) were detected as being positively selected with a pp > 0.99 (Table S4). These, however, are different sites from the ones predicted by model M2a to be positively selected, thus bringing the combined total number of sites to 11. For *P. omphalodes*, the branch-site model tests were able to detect three branches under putative positive selection, including one gene and two internal branches (Fig. S11, Table S4). Additionally, two of the three branches remained significant after correcting for multiple testing (*P < 0.05*), while two residues (22Y and 23V) were detected as being positively selected with pp > 0.99.

Overall, the combination of the site models and branch-site models of *PAML* could perhaps support a model where paralogous CBM14-containing proteins diversify early in their evolution soon after the episodic bursts of gene duplication in some of the lineages, while their rate of evolution considerably decelerates during later stages to possibly maintain functional diversity among them. It should be noted that since three to four CBM14s were likely to have been present in the LOCA, it is possible that the functional diversification took place already at that stage. However, the fact that our branch-site models still detect branches under positive selection in both *D. melanogaster* and *P. omphalodes* indicates that there is still ongoing evolution in the CBM14 family.

Finally, to further understand the functional association of the putatively positively selected residues in the CBM14s of *D. melanogaster* and *P. omphalodes*, we mapped these residues to the solution structure of tachycitin, a 73 amino acid lectin from the horseshoe crab *T. tridentatus* with a single CBM14 in its sequence and currently the only known structural representative of the CBM14 family [7,8]. The C terminus of the CBM14, and especially the region from C40 to G60, was predicted to form the putative ligand-binding pocket, and therefore play a role in chitin-binding activity [8]. In *D. melanogaster*, the positively selected residues P29 and H38 aligned to the P41 and L50 amino acids in tachycitin, respectively, thus mapping within the predicted chitin-binding region of CBM14 (Fig. S12). The presence of positively selected sites within the substrate-binding pocket of CBM14 could suggest an effect on the functional properties, by altering binding affinity for chitin and/or specificity for different carbohydrates. The remaining nine selected residues in *D. melanogaster* and the two selected residues in *P. omphalodes* are all located in the N-terminal region, and four of these residues are clustered within the beta-strands of the CBM14. In contrast to the C-terminal region, the N-terminal region of the CBM14 is responsible for the stability of the protein structure and for the correct positioning of the C-terminal chitin-binding pocket towards its substrate [8]. Therefore, the selected residues at the N terminus could be involved in fine-tuning CBM14 stability and conformation.

### Concluding remarks

In this study, we reveal the global, complex and fascinating evolutionary history of a dynamic CBM family of high interest that can be used as a reference for other CBM families as well. Next to the sequences deposited in the CAZy database, our comprehensive genome-wide searches uncovered an impressive number of CBM14 sequences from diverse sets of phylogenetic lineages across almost all domains in the tree of life. The taxonomic distribution of this family predominantly within the animal and fungal kingdoms, and its highly disseminated presence in other eukaryotic lineages and domains of life, suggest a later evolutionary origin and a high rate of inheritance through HGT. Further in-depth analyses reveal that the CBM14 family underwent significant sequence diversification within its evolutionary history, coupled with multiple episodic lineage-specific expansions and contractions that probably reflect adaptations to particular lifestyles. In many lineages, bursts of duplication in genes encoding CBM14-containing proteins, many of which bear the signatures of adaptive evolution, support a scenario of functional diversification in early stages after duplication. Taken together, these results highlight the importance of CBMs as fundamental building blocks during protein evolution and shed light on their remarkable evolutionary dynamics.

### Materials and methods

#### Retrieval of CBM14 sequences from species across the tree of life

CBM14 sequences were collected using an initial set of 12 fungal, 56 insectal and 10 bacterial CBM14 sequences that have been previously characterized in the literature (Table S5) as queries to search against the public protein/nucleotide/genome databases (Fig. S13). Retrieved sequences with an e-value < 1E-5 were considered as true homologs. Profile hidden Markov models (HMMs) were also constructed to identify and retrieve more distantly related members of the CBM14 family. The HMM profiles deposited in Pfam [49] were used to identify the CBM14s and additional
domain types present in CBM14 proteins. The final collection of sequences was cross-validated with the CBM14 data from Pfam, as well as by aligning the sequences back to the respective genomes in order to confirm the presence of CBM14s. To examine the levels of homology among CBM14 sequences, pairwise similarity scores were calculated by sequence shuffling [50]. The produced pairwise e-values were used for hierarchical clustering. The motifs were identified by a combination of PURGE and GLAM2 from the MEME suite [51].

Phylogenetic reconstruction of the CBM14 family and inference of horizontal gene transfer (HGT)

A phylogenetic tree was constructed with all the CBM14 family members using fastTree [29] and clades with strong local support values containing CBM14 sequences from Opisthokonta and non-Opisthokonta species were examined for the presence of HGT using the statistical tests implemented in CONSEL [31] and a maximum statistical agreement forest algorithm implemented in PRUNIER [33]. HGT was inferred when all the test results consistently supported HGT. The phylogenetic independent contrasts approach [52,53] was applied to infer the ancestral CBM14 copy number. For this purpose, a reference species tree was built with data from the sTOL database [29] and taxonomy information from NCBI. We next modeled the number of CBM14s per species as a continuous evolutionary trait using Brownian motion. Gain and loss of CBM14s was inferred by comparing the putative CBM14 copy number of each node against that of the respective parent node.

Ancestral state reconstruction and inference of gain and loss events within Opisthokonta

To infer the ancestral domain copy number and gain and loss events within Opisthokonta, we applied a reconstruction method based upon a least-squares approach (phylogenetic independent contrasts) [52,53]. The reference species tree phylogeny for the copy number inference was built with data obtained from the sTOL database [29] and taxonomy information from NCBI (Fig. S5). Given the highly skewed CBM14 copy number in different lineages and extant species, modeling the copy number variation with a Bayesian method or as a discrete variable was computationally intractable. Therefore, we instead modeled the counts of CBM14s in each species as a continuous evolutionary trait following a Brownian motion model, i.e. modeling the number of CBM14s that have evolved randomly in any direction [53]. In this model, the difference in CBM14 counts between any two given species has a distribution of the CBM14 number centered on zero and a variance in proportion to the divergence time. The same calculation was also performed for the internal nodes to reconstruct the ancestral states. Gain and loss of CBM14s in nodes were inferred by comparing the putative CBM14 copy number in each node against the copy number of the respective parent node. In cases with decreased values from parent to child nodes a loss event was inferred, while in cases with increased values from parent to child nodes a gain event was inferred. Custom scripts were coded in PERL and R to perform the above analyses. The final result was visualized by a python script using the ete2 modules [54].

Detection of adaptive evolution in paralogous CBM14-containing proteins

We investigated instances of departure from the neutral model of molecular evolution in the CBM14 family by analyzing a set of paralogous CBM14-containing proteins from the common fruit fly D. melanogaster and the Ascomycete P. omphalodes. In the analysis, only proteins with a single CBM14 and no additional domains other than the signal peptide were considered in order to minimize false inference of positive selection due to the tandem CBM14 duplication or adaptive convergence between the different domains present in modular CBM14 proteins. CBM14 amino acid sequences were initially aligned using PRANK [55] incorporated with MAFFT [56] and EXONERATE [57] to accelerate the alignment process, and were then back-translated into nucleotide alignments. The alignment process was iterated five times in order to identify the best alignment, which was subsequently inspected manually to assure the quality of the final alignment. TRIMAL was used to remove the codon sites with more than 75% gaps in the alignment, while pairwise Ks analysis was also applied in order to remove highly divergent (Ks > 2) or highly similar (Ks < 0.02) sequences that could impair detection of positive selection [58]. Next, the Genetic Algorithm Recombination Detection (GARD) algorithm of the HYPHY package was used to confirm that there were no recombination breakpoints in our data set that could lead to false detection of positive selection and bias results [59]. RAxML [60] was used for ML tree construction based on the PROTGAMMA model with 1000 bootstraps. Evidence of adaptive evolution in the sequences was investigated by means of LRTs that took into account the ratio (ω) of non-synonymous (dN) to synonymous (dS) nucleotide substitution rates. Both site and branch-site models as implemented in the CODEML program of the PAML software package (v4.4) were used to investigate changes in evolutionary pressure in the selected members of the CBM14 family that could be indicative of functional diversification [43,61]. Calculation of likelihood values was performed three times in all cases for both the alternative and null hypotheses of each test in order to avoid potential convergence problems [62]. Sites under positive selection were subsequently identified based on the BEB posterior probability. A cutoff of 0.99 of BEB
probability was used for inference of sites under positive selection. Finally, the selected sites were plotted to the CBM14 structure using the chimera program [63].

Acknowledgements
This work used resources of the UC Davis CAES Computing Center and was financially supported by UC Davis faculty start-up funds. We kindly acknowledge Dr Antonis Rokas for useful suggestions and discussions. Dr Neil McRoberts, Dr Amanda Kohler and Anthony Salvucci are kindly acknowledged for critically reading the manuscript.

Author contributions
TC: performed analyses; analyzed data; wrote the paper. IS: planned analyses; analyzed data; wrote the paper.

References


Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website:

Fig. S1. Hierarchical clustering analysis of the identified CBM14 family members displayed as a heatmap.

Fig. S2. Motifs present in the five major clustering groups formed by CBM14 family members, as these were defined based on the hierarchical clustering analysis.

Fig. S3. Phylogenetic analysis of the entire CBM14 family.

Fig. S4. Putative cases of eukaryote-to-eukaryote horizontal gene transfer (HGT) in the CBM14 family, involving the transfer of sequences to species of Viridiplantae (green plants), Stramenopiles (Chromalveolata) and Amoebozoa.

Fig. S5. Ancestral state reconstruction and inference of gain and loss events in the evolutionary history of the CBM14 family within Opisthokonta.

Fig. S6. Boxplots comparing the total number of CBM14s recovered from the major phylogenetic lineages in the tree of life.

Fig. S7. Boxplots comparing the total number of CBM14-containing proteins recovered from the major phylogenetic lineages in the tree of life.

Fig. S8. Correlation between the number of CBM14s and the number of CBM14-containing proteins in lineages where a large expansion of the CBM14 family has taken place, including (A) all taxa, (B) Nematoda, (C) Drosophila, (D) Culicidae, (E) Aporcita and (F) Tunicata.

Fig. S9. The distribution of the number of CBM14s present in CBM14-containing proteins.

Fig. S10. Phylogeny of the CBM14 family members obtained from proteins in Drosophila melanogaster with a single CBM14 and no other domains than the signal peptide that was used to determine positive selection by the branch-site models implemented in PAML.

Fig. S11. Phylogeny of the CBM14 family members obtained from proteins in Pyronema omphalodes with a single CBM14 and no other domains than the signal peptide that was used to determine positive selection by the branch-site models implemented in PAML.

Fig. S12. Mapping of amino acid residues under positive selection on the solution structure of CBM14.

Fig. S13. Flowchart of the searches employed in order to retrieve CBM14 sequences from public sequence databases, and of the downstream molecular evolutionary analyses.

Table S1. Sequences for examining the HGT events.

Table S2. Log-likelihood values and parameter estimates under the site models implemented in CODEML for the selected CBM14 sequences from Pyronema omphalodes.

Table S3. Log-likelihood values and parameter estimates under the branch-site models implemented in CODEML for the selected CBM14 sequences from Drosophila melanogaster.

Table S4. Log-likelihood values and parameter estimates under the branch-site models implemented in CODEML for the selected CBM14 sequences from Pyronema omphalodes.

Table S5. Seed sequences for retrieval of homologous CBM14 sequences.

Doc. S1. List and characteristics of the CBM14 members recovered from species across the tree of life.