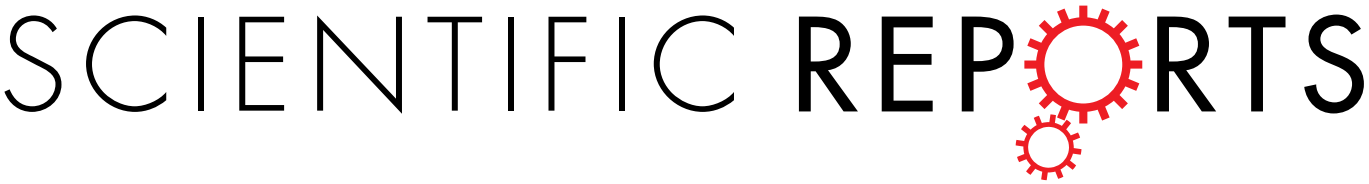

OPEN

Identification of a novel family of carbohydrate-binding modules with broad ligand specificity

Received: 18 June 2015
Accepted: 03 November 2015
Published: 14 January 2016

Cheng-Jie Duan, Yu-Liang Feng, Qi-Long Cao, Ming-Yue Huang & Jia-Xun Feng

Most enzymes that act on carbohydrates include non-catalytic carbohydrate-binding modules (CBMs) that recognize and target carbohydrates. CBMs bring their appended catalytic modules into close proximity with the target substrate and increase the hydrolytic rate of enzymes acting on insoluble substrates. We previously identified a novel CBM ($CBM_{C5614\text{-}1}$) at the C-terminus of endoglucanase C5614-1 from an uncultured microorganism present in buffalo rumen. In the present study, that the functional region of $CBM_{C5614\text{-}1}$ involved in ligand binding was localized to 134 amino acids. Two representative homologs of $CBM_{C5614\text{-}1}$, sharing the same ligand binding profile, targeted a range of β-linked polysaccharides that adopt very different conformations. Targeted substrates included soluble and insoluble cellulose, β-1,3/1,4-mixed linked glucans, xylan, and mannan. Mutagenesis revealed that three conserved aromatic residues (Trp-380, Tyr-411, and Trp-423) play an important role in ligand recognition and targeting. These results suggest that $CBM_{C5614\text{-}1}$ and its homologs form a novel CBM family (CBM72) with a broad ligand-binding specificity. CBM72 members can provide new insight into CBM-ligand interactions and may have potential in protein engineering and biocatalysis.

Most carbohydrate-active enzymes are modular proteins that comprise two or more discrete catalytic modules (CMs) and non-catalytic carbohydrate-binding modules (CBMs) connected by linker sequences[1,2]. In the most recent update of the Carbohydrate-Active enZYmes database, CBMs were classified into 71 families based on amino acid sequence similarity (http://www.cazy.org)[3]. Some CBM families are classified into subfamilies based on key residues in the ligand binding site (e.g. CBM3[4]) or topological structure of the ligand binding region (e.g. CBM2[5]). However, an alternative classification based on the structure of the ligand binding site grouped these protein modules into three types: surface-binding, glycan-chain-binding, and small sugar-binding (Types A–C), respectively[1]. Recently, Gilbert *et al.* (2013) expanded on this classification and proposed that Type A enzymes recognize the surface of crystalline polysaccharides, Type B bind the internal regions of glycan chains (endo-type), and Type C bind the termini of glycans (exo-type)[6].

The main function of CBMs is to recognize and bind polysaccharides and to enhance the hydrolytic activity of the appended hydrolase against insoluble substrates by increasing the effective enzyme concentration at the substrate surface[1,6,7]. Aromatic amino acids within the ligand binding site of CBMs play an important role in ligand recognition through hydrophobic interactions[1,8]. Some CBMs disrupt the ordered structure of recalcitrant substrates and increase their accessibility to the appended enzymes[9–12].

Novel CBM families are frequently discovered and added to the appropriate databases. Recently, CBM66 was found to confer exolevanase activity on non-specific fructosidase through an avidity mechanism in which the CBM and CM target the termini of different branches of the same polysaccharide molecules[13]. The binding of expansin EXLX1, a representative member of CBM63, to whole cell walls is mediated by electrostatic and polar interactions between the basic D2 domain and the acidic polysaccharide matrix[14]. However, the crystal structure of the EXLX1-cellohexaose complex revealed that one cellohexaose molecule is packed between the aromatic surfaces of the D2 domains of two neighbouring EXLX1 molecules in a unique protein:ligand arrangement[15]. Identification of other novel family of CBMs may reveal further novel CBM-ligand interaction mechanisms.

Metagenomic approaches have been widely applied to the discovery of novel biocatalysts from diverse environmental samples[16]. CBM59, which binds efficiently to mannan, xylan and cellulose, was identified from

State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, The Key Laboratory of Ministry of Education for Microbial and Plant Genetic Engineering, and College of Life Science and Technology, Guangxi University, 100 Daxue Road, Nanning, Guangxi, 530004, China. Correspondence and requests for materials should be addressed to C.-J.D. (email: cjduan@gxu.edu.cn) or J.-X.F. (email: jiaxunfeng@sohu.com)

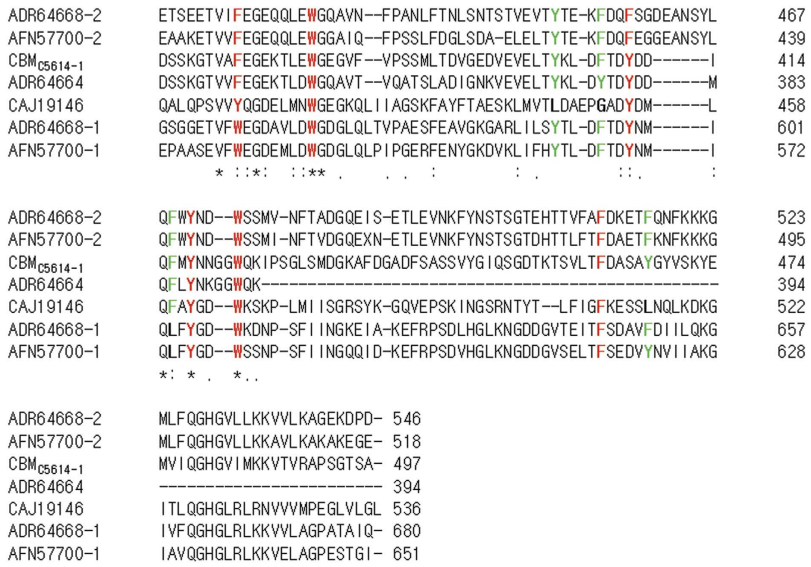

Figure 1. Multiple sequence alignment of $CBM_{C5614-1}$ and selected homologs. Sequence identity and similarity are indicated by asterisks and dots, respectively. Completely conserved aromatic amino acid residues are red, and partially conserved aromatic amino acid residues are green. Numbers behind the each alignment represent the position of amino acids in the source enzyme of each member of the novel carbohydrate binding module family.

metagenomes of cow manure samples[17], and a novel starch-binding domain belonging to CBM69 was identified from the α-amylase derived from a marine metagenomic library[18].

In our previous work, we identified a novel CBM from the C-terminus of endoglucanase C5614-1 (GenBank: ACA61140) derived from an uncultured microorganism present in buffalo rumen. $CBM_{C5614-1}$ is composed of 189 amino acids, shares no significant homology to known CBMs, and can bind to a broad range of soluble polysaccharides such as barley glucan, methylcellulose, 2-hydroxyethycellulose, birch wood xylan, and carboxymethyl cellulose (CMC), as well as insoluble polysaccharides such as Avicel, acid swollen cellulose (ASC) and lichenan[19].

In the present study, $CBM_{C5614-1}$ and its homologs were found to constitute a novel family of CBMs (CBM72) with a broad ligand-binding specificity. Three key conserved aromatic residues in the ligand binding site were identified and characterized.

Results

Residues 364-497 of $CBM_{C5614-1}$ constitute the polysaccharide binding region. In our previous work, we identified a novel CBM (189 aa) at the C-terminus (residues 349–537) of endoglucanase C5614-1 derived from an uncultured microorganism present in buffalo rumen[19]. Analysis of the isoelectric point (pI) of $CBM_{C5614-1}$ with DNAStar revealed that the 40 aa basic C-terminal domain (BTD) has a pI of 9.58, while the pI of the remaining 149 aa is 4.5. In order to determine the shortest functional region capable of ligand binding, N- and C-terminal truncations of $CBM_{C5614-1}$ were constructed. Binding experiments of deletion mutants showed that deletion of the C-terminal BTD did not affect the binding capacity towards soluble or insoluble polysaccharides, and neither did removing 15 aa from the N-terminus. However, removing 30 aa from the N-terminus resulted in a complete loss of binding to soluble polysaccharides. Residues 364–497 were therefore confirmed to house the ligand-binding region of $CBM_{C5614-1}$ (Supplementary Fig. S1). Unless otherwise stated, all subsequent mention of $CBM_{C5614-1}$ below refers to residues 364–497 of endoglucanase C5614-1.

Sequence analysis of $CBM_{C5614-1}$ homologs. A BLAST search using residues 364–497 identified seven homologous peptides in the NCBI protein database sharing 28–99% identity and 48–99% similarity with $CBM_{C5614-1}$. All homologous peptides were found to be located at the C-terminus of glycoside hydrolases from uncultured microorganisms present in the rumen[20–23]. Of the seven homologs, AEK98797 (GeneBank number) is the most closely related, with only a single amino acid difference. ADR64668 and AFN57700 contained two tandem homologous regions, which were named ADR64668-1(549-680), ADR64668-2(411–546), AFN57700-1(520–651) and AFN57700-2(384–518), respectively. Unlike the other homologs, ADR64664 has an incomplete C-terminus.

Aromatic amino acids within the ligand binding site are known to play an important role in recognizing and binding polysaccharides in all CBMs[1]. Sequence alignment revealed six aromatic amino acids that are completely conserved in all homologs, including Phe-372, Trp-380, Tyr-411, Tyr-418, Trp-423 and Phe-462 (numbered according to source enzyme C5614-1 of $CBM_{C5614-1}$) (Fig. 1, red). Four partially conserved aromatic amino acids were also identified (Tyr-404, Phe-408, Phe-416 and Tyr-467) (Fig. 1, green).

Representative $CBM_{C5614-1}$ homologs also exhibited broad ligand specificity. ADR64668-1(549–680) and CAJ19146(405–536) that share 35% and 28% sequence identity with $CBM_{C5614-1}$, respectively, were

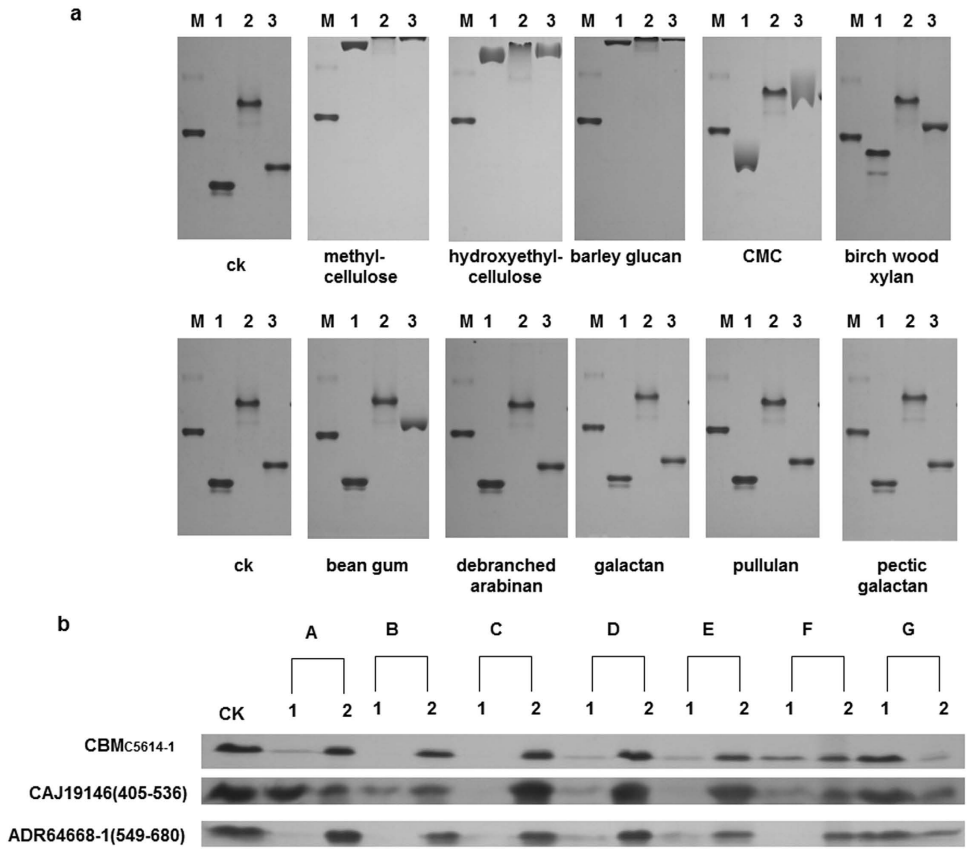

Figure 2. Binding of $CBM_{C5614-1}$ and its homologs to polysaccharides. (**a**) Binding of $CBM_{C5614-1}$ and its homologs to soluble polysaccharides. Proteins and BSA were separated using non-denaturing polyacrylamide gels containing 0.1% (wt/vol) soluble polysaccharides. A gel without polysaccharides (CK) served as a control. M: BSA control (no polysaccharides bound); 1: $CBM_{C5614-1}$; 2: CAJ19146 (405–536); 3: ADR64668-1 (549–680). (**b**) Binding of $CBM_{C5614-1}$ and its homologs to insoluble polysaccharides. 30 μg of purified $CBM_{C5614-1}$ and its homologs were incubated with 200 μl 4% (wt/vol) insoluble polysaccharide including Avicel (**A**), ASC (**B**), insoluble birch wood xylan (**C**), mannan **(D)**, lichenan (**E**), raw starch from cassava (**F**) or agarose (**G**). The same amount of protein used in the binding assay but without polysaccharide was included as a control (CK). After centrifugation, unbound protein in the supernatant (lane 1) and bound proteins in the precipitate (lane 2) were analyzed by SDS-PAGE.

expressed in *Escherichia coli (E. coli)*, and the ligand binding capacity of the recombinant proteins was compared with that of $CBM_{C5614-1}$. The homologs exhibited the same broad polysaccharide binding profiles (Fig. 2). Affinity gel electrophoresis revealed extensive binding to methyl cellulose, hydroxyethyl cellulose and barley glucan, weak binding to birch wood xylan, CMC and bean gum, and no binding to de-branched arabinan, galactan, pullulan and pectic galactan. Only $CBM_{ADR64668-1}$ showed significant binding to CMC and bean gum (Fig. 2a). $CBM_{C5614-1}$ and both homologs could bind to the insoluble polysaccharides Avicel, ASC, mannan, lichenan, birch wood xylan, and raw starch from cassava, but not to agarose (Fig. 2b). These results confirmed that the homologs were indeed CBMs.

Three conserved aromatic amino acids play an important role in ligand binding in CBM72 proteins. The 10 conserved aromatic amino acids of $CBM_{C5614-1}$ were substituted with alanine using an overlap PCR method, mutant constructs were expressed in *E. coli*, and recombinant proteins were tested for binding to polysaccharides. Affinity gel electrophoresis revealed that $CBM_{C5614-1}$ mutants W380A and W423A completely lost their ability to bind soluble polysaccharides, the affinity of Y411A was also diminished, however the other seven mutants displayed only a slight decrease in affinity or none at all, compared with wild type $CBM_{C5614-1}$ (Fig. 3a).

The binding capacity of W380A and W423A to insoluble polysaccharides decreased dramatically and that of Y411A also decreased to some degree (Fig. 3b). In contrast, the other seven mutants had no obvious change compared with wild type $CBM_{C5614-1}$. Trp-380, Tyr-411 and Trp-423 are therefore important for polysaccharide binding.

Studies by Sunna *et al.*[24] and Hachem *et al.*[25] suggested that 25% sequence identity is an appropriate threshold for CBM family members. $CBM_{C5614-1}$ and its homologs all share more than 25% sequence identity with each other. The three key aromatic amino acids involved in substrate binding, are well conserved among all the homologs. These proteins therefore form a novel family of CBMs that we propose to call the CBM72 family. The

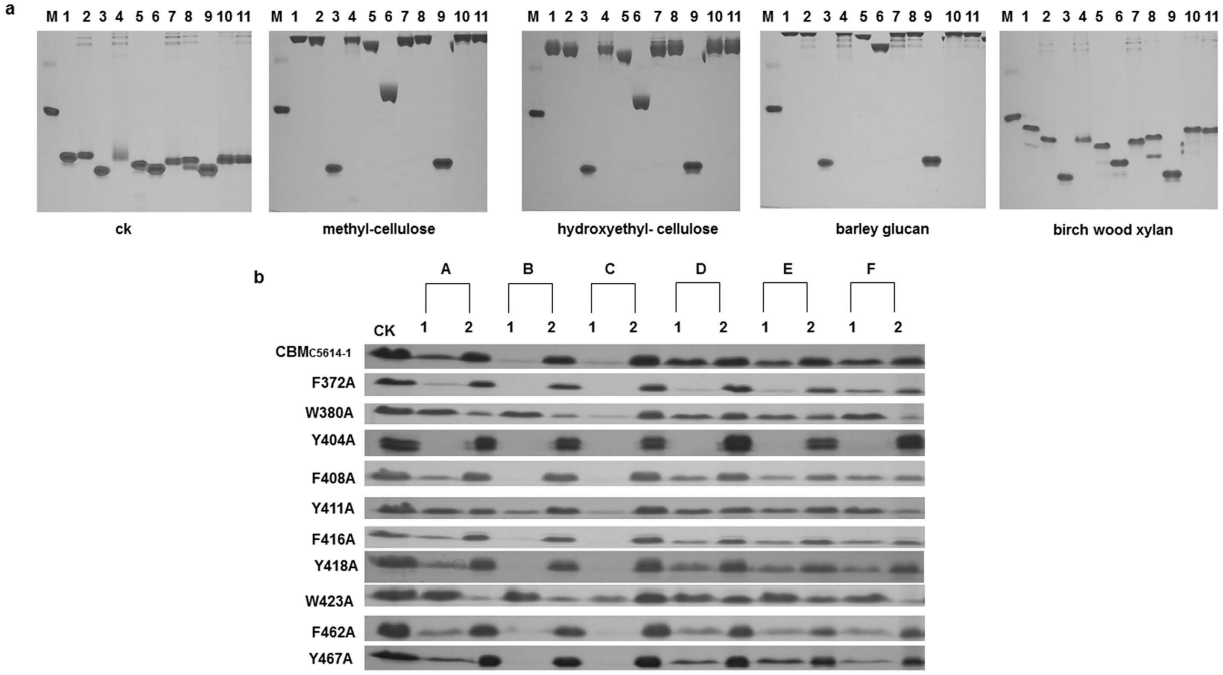

Figure 3. Binding of $CBM_{C5614-1}$ variants to polysaccharides. (**a**) Binding of $CBM_{C5614-1}$ variants to soluble polysaccharides. Proteins and BSA were separated using non-denaturing polyacrylamide gels containing 0.1% (wt/vol) soluble polysaccharides. A gel without polysaccharides served as a control (CK). M: BSA non-binding control; 1: $CBM_{C5614-1}$; 2: F372A; 3: W380A; 4: Y404A; 5: F408A; 6: Y411A; 7: F416A; 8: Y418A; 9: W423A; 10: F462A; 11: Y467A. (**b**) Binding of $CBM_{C5614-1}$ variants to insoluble polysaccharides. 30 μg of purified protein was incubated with 200 μl 4% (wt/vol) insoluble polysaccharide including Avicel (**A**), ASC (**B**), insoluble birch wood xylan (**C**), mannan (**D**), lichenan (**E**) and raw starch from cassava (**F**). The same amount of protein used in the binding assay but without polysaccharide was included as a control (CK). After centrifugation, unbound protein in the supernatant (lane 1) and bound proteins in the precipitate (lane 2) were analyzed by SDS-PAGE.

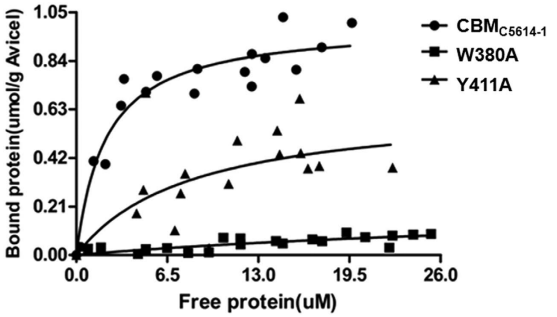

Figure 4. Depletion isotherms of wild type $CBM_{C5614-1}$ and its mutants binding to Avicel. Binding isotherms were carried out as described in Materials and Methods.

broad substrate binding specificity suggests that CBM72 members belong to type B (endo-type) CBMs that bind internally to glycan chains[6].

Quantitative binding of $CBM_{C5614-1}$ and its mutants to insoluble cellulose. The dissociation constant (K_d) of wild type $CBM_{C5614-1}$ binding to Avicel was 2.1, which was comparable to K_d values of other CBMs including CBM10, CBM3, CBM63, CBM17 and CBM28[15]. Affinity for Avicel was too low to determine the K_d of mutant W423A, but W380A exhibited a 20-fold decrease in affinity, and Y411A displayed a 4-fold reduction, compared to wild type $CBM_{C5614-1}$ (Fig. 4; Table 1).

Secondary structure determination of $CBM_{C5614-1}$ and its mutants by CD spectroscopy. The secondary structure of $CBM_{C5614-1}$ and its mutants was determined by CD spectroscopy. The composition of the secondary structures of mutants W380A, Y411A and W423A was only slightly different to that of wild type $CBM_{C5614-1}$ (Supplementary Fig. S2; Supplementary Table S2), indicating that these conserved aromatic residues do not contribute to maintaining the protein structure and may be the key amino acids in the ligand binding site.

Protein	K_d	B_{max}
	μM	μmol/g Avicel
$CBMC_{C5614-1}$	2.10	1.00
W380A	43.67	0.23
Y411A	8.22	0.65
W423A	[a]	[a]

Table 1. Affinity of wild type and variants of $CBM_{C5614-1}$ for Avicel as determined by depletion isotherms. [a]Too low to be determined accurately.

$CBM_{C5614-1}$ contains 44% β-strands structure. This high proportion of β-strands is consistent with other known CBMs[1].

Discussion

This study describes a novel CBM family, CBM72, members of which that display a particularly broad ligand binding specificity and target a range of β-linked soluble and insoluble polysaccharides that adopt very different conformations, including cellulose, β-1,3/1,4-mixed linked glucans, xylan, and mannan. Generally, CBMs that bind β-glucan chains often display broad specificity and recognize β-1,4-glucans (cellulose), β-1,3/1,4-mixed linked glucans, xyloglucan and other β-1,4-glycans, examples of which include CBM family 6[26], 60[27], 62[28], and 65[29,30]. CBM72 proteins can bind to mannans (O3-C3-C2-O2 torsion angle of mannose = −60°) and β-1,4-glucans and -xylan (O3-C3-C2-O2 torsion angle of monosacccharide = + 60°)[27], whereas other CBMs such as CBM60 can target the key geometric signature of β-1,4-glucans and β-1,4 xylan, but are not capable of binding to mannan[27]. This indicates that the binding site of CBM72 members may possess greater plasticity in order to accommodate very different ligand conformations.

The representative source enzyme endoglucanase C5614-1 of CBM72 showed the same broad substrate specificity observed previously[19]. C5614-1 can hydrolyze different β-1,4-linked polysaccharides but is most active towards barley glucan, followed by CMC, lichenan, 2-hydroxyethyl cellulose, methyl cellulose and xylan[19]. Generally, CBMs do not increase the hydrolytic activity of appended catalytic modules towards soluble substrates[31,32]. In some cases, the CBM truncated mutant Egl330 showed improved turnover rate (k_{cat}) and catalytic efficiency (k_{cat}/K_m) with CMC as substrate compared with wild type Egl499[33]. In our previous study, deletion of $CBM_{C5614-1}$ from endoglucanase C5614-1 increased its hydrolytic activity against barley glucan[19], despite the high affinity of $CBM_{C5614-1}$ for this substrate, suggesting that this CBM hindered its catalytic module to degradation of barley glucan. This result indicates that there may be no correlation between hydrolytic efficiency of the appended catalytic module and binding capacity of the CBM towards soluble substrates.

We next investigated how CBM72 is able to bind a broad range of ligands, and considered the possibility of more than one binding sites on the surface for this purpose. CBMs in family 6 were reported to include three ligand binding sites with distinct specificities, and these proteins also bind a broad ligand range including β1,4-1,3-mixed linked glucans, cello-oligosaccharides, insoluble forms of cellulose, the β1,3-glucan laminarin, and xylooligosaccharides[26,34,35]. Mutagenesis of the conserved aromatic amino acids in $CBM_{C5614-1}$ showed that Trp380 and Trp423 were crucial for binding barley glucan, 2-hydroxyethyl cellulose, birch wood xylan and Avicel, suggesting that all these diverse ligands were bound to the same binding site of $CBM_{C5614-1}$. The basis for the broad binding specificity remains unknown without structural data, and determining the structure of a CBM72-ligand complex is a priority in future studies.

The biological function of the unusually broad ligand specificity of CBM72 proteins also remains unknown, but targeting and proximity effects of CBMs may play an important role. Herve *et al.* (2010) demonstrated that CBMs can potentiate the action of appended catalytic modules toward polysaccharides in intact cell walls through the recognition of non-substrate polysaccharides, provided that the non-substrate and substrate are close to each other. For example, the capacity of xylanases to degrade xylan in secondary walls was potentiated by both xylan and cellulose-directed CBMs[36]. The plant cell wall is highly complex and is comprised of various polysaccharides such as cellulose, pectin and hemicellulose that crosslink with each other to form an intricate meshwork[37]. Catalytic modules attached to CBMs with broad ligand specificity may be easier to target to their intended substrates than those appended to CBMs with strict ligand specificity.

In conclusion, this study reports a new family of CBMs with a uniquely broad ligand binding specificity. Three conserved CBM aromatic amino acids were found to play an important role in recognizing and targeting ligands. The presence of CBM72 domains may facilitate the targeting of catalytic domains to their substrates in plant cell walls. These novel CBMs can provide new insight into CBM-ligand interactions and may have potential for engineering of cellulase or other enzymes for improved biotechnological performance.

Materials and Methods

Sequence analysis. Protein modular structure was predicted using SMART (http://smart.embl-heidelberg.de) and Pfam (http://pfam.xfam.org/search/sequence). Multiple alignments of $CBM_{C5614-1}$ and its homologs were performed using ClustalW (http://www.ebi.ac.uk/Tools/ClustalW).

Sources of carbohydrates. Carboxymethyl cellulose, birch wood xylan, hydroxyethyl cellulose, methyl cellulose, bean gum and Avicel were purchased from Sigma (St. Louis, MO). Cello-oligosaccharides, barley glucan,

galactan, debranched arabinan, pullulan, pecticgalactan and lichenin were purchased from Megazyme international Ireland Ltd (Bray, Ireland).

Cloning and expression of wild type and mutant $CBM_{C5614-1}$ and its homologs. The cosmid C5614-1 (GenBank: EU449484), derived from metagenomics experiments on ruminal microbiomes[38], were used as templates for amplifying the gene encoding $CBM_{C5614-1}$. DNA fragments were amplified using Primestar polymerase (TaKaRa, Kyoto, Japan) and an appropriate combination of primers (Supplementary Table S1) that included NdeI or XhoI restriction sites (underlined), and amplified products were purified, digested with NdeI and XhoI, and cloned into the expression vector pET-30a(+) (Novagen, Darmstadt, Germany) in order to express recombinant proteins with 6 × His tags at the C-terminus.

Genes encoding $CBM_{ADR64668-1}$ (549–680) and $CBM_{CAJ19146}$ (405–536) were synthesized artificially based on the DNA sequences of $CBM_{C5614-1}$ homologs (GenBank: ADR64668, CAJ19146) following codon optimization to maximize expression in *E. coli*, and ligated into the NdeI and XhoI restriction sites of expression vector pET-30a(+).

Constructs encoding deletion derivatives of $CBM_{C5614-1}$ (349–537) were constructed by deleting the nucleotides encoding successively the N-terminal 15 aa, and the C-terminal 40 aa, using PCR. The following constructs were prepared: $CBM_{C5614-1}$ (349–497), $CBM_{C5614-1}$ (364–497) and $CBM_{C5614-1}$ (379–497), where numbers in brackets correspond to the amino acid sequence of endoglucanase C5614-1. PCR products were ligated into the expression vector pET-30a(+). Primers used for PCR amplification are shown in Supplementary Table S1.

Variants of $CBM_{C5614-1}$ were engineered by site-directed mutagenesis using an overlap-extension PCR procedure based on that described by Ho *et al.*[39] using primers shown in Supplementary Table S1.

Purification of recombinant proteins. All constructs were verified by DNA sequencing and transformed into *E. coli* Rosetta (DE3) pLysS (Novagen) for protein expression and positive clones were selected by growth on kanamycin- and chloramphenicol-containing media. Cells harboring the recombinant plasmids were grown to an OD_{600} of 0.6 in LB broth containing 25 μg/mL kanamycin and 34 μg/mL chloramphenicol at 37 °C with shaking at 200 rpm. The expression of the target genes was induced by adding 0.5 mM isopropyl-β-D -1-thiogalactopyranoside to the medium and continuing the incubation at 20 °C with shaking at 100 rpm overnight. Recombinant proteins were extracted from the cytoplasmic fraction of cell lysates and purified by affinity chromatography with Nickel-nitrilotriacetic acid agarose resin (Ni-NTA, Qiagen) according to the manufacturer's instructions. Purified proteins were desalted using Amicon Ultra-10 ultrafiltration columns (Millipore, Billerica, MA) and diluted into citrate/phosphate buffer, pH 5 (a mixture of 100 mM citric acid and 200 mM Na_2HPO_4 at a volume ratio of 97:103).

Protein determination. Protein concentration was determined using the Micro BCA kit (Pierce, Rockford, IL) with bovine serum albumin (BSA) as the standard. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was performed on 12% polyacrylamide gels using the method of Laemmli[40] and proteins were visualized by Coomassie Blue staining.

Soluble polysaccharide binding assays. The capacity of $CBM_{C5614-1}$, its homologs and variants to bind to soluble polysaccharides was determined by affinity gel electrophoresis performed as described by Duan *et al.*[19]. Polysaccharides were incorporated into the gel at a concentration of 0.1% prior to polymerization. A control gel without polysaccharides was prepared and run simultaneously. Electrophoresis was conducted at 100 V for 4 h at 4 °C. BSA that does not bind polysaccharides was used as a negative control.

Insoluble polysaccharide binding assays. Proteins (30 μg) were mixed with 4% (wt/vol) insoluble polysaccharides in 0.2 mL citrate/phosphate buffer (pH 5) and incubated on ice for 5 h with occasional stirring. After centrifugation at 10,000 g, 4 °C for 10 mins, supernatants (unbound proteins) were collected and pellets were washed twice with 1 mL citrate/phosphate buffer (pH 5). Polysaccharides bound to proteins were then eluted with 100 μL 2% SDS for 30 min at 37 °C. Eluted proteins were collected by centrifugation and subjected to SDS-PAGE. Controls with proteins but no ligands were included to ensure that precipitation did not occur during the assay.

Depletion isotherms to quantify the binding of wild type $CBM_{C5614-1}$ and its variants to Avicel were carried out by mixing protein 1–100 μM protein with 0.2 mL citrate/phosphate buffer (pH 5) containing 1% Avicel. The mixture was shaken on a table concentrator (TENSUG) at 340 rpm, 4 °C until equilibrium was reached (5 h). Samples were centrifuged at 10,000 g, 4 °C for 2 min to pellet the bound substrate, and unbound proteins in the supernatant were quantified using the Pierce BCA protein assay kit using the formula: bound protein = total protein—unbound protein. Dissociation constants (K_d) and Bmax values (amount of protein bound at saturation) were calculated by fitting the data to a single site Langmuir isotherm using Graphpad Prism 5 (GraphPad Software, Inc., San Diego, CA). At least three separate binding isotherms were carried out for each protein.

Circular Dichroism (CD) Spectroscopy. Proteins were dialyzed extensively against 5 mM sodium phosphate (pH 5) and CD spectra were collected on a Biologic MOS-450 spectropolarimeter between 188–250 nm with steps of 60 nm per minute. All samples were analyzed in triplicate. After subtraction of buffer spectra, CD spectra were smoothed using the means-movement method and analyzed using CDROM Biokine 4 and Origin 8.0 that utilize the K2D method[41] to analyze protein secondary structure.

References

1. Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J.* **382,** 769–781 (2004).
2. Shoseyov, O., Shani, Z. & Levy, I. Carbohydrate binding modules: biochemical properties and novel applications. *Microbiol Mol Biol Rev* **70,** 283–295 (2006).

3. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42,** D490–495 (2014).
4. Cai, S., Zheng, X. & Dong, X. CBM3d, a novel subfamily of family 3 carbohydrate-binding modules identified in Cel48A exoglucanase of *Cellulosilyticum ruminicola*. *J Bacteriol* **193,** 5199–5206 (2011).
5. Simpson, P. J., Xie, H., Bolam, D. N., Gilbert, H. J. & Williamson, M. P. The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J Biol Chem* **275,** 41137–41142 (2000).
6. Gilbert, H. J., Knox, J. P. & Boraston, A. B. Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin Struct Biol* **23,** 669–677 (2013).
7. Beguin, P. & Aubert, J. P. The biological degradation of cellulose. *FEMS Microbiol Rev* **13,** 25–58 (1994).
8. Guillen, D., Sanchez, S. & Rodriguez-Sanoja, R. Carbohydrate-binding domains: multiplicity of biological roles. *Appl Microbiol Biotechnol* **85,** 1241–1249 (2010).
9. Din, N. *et al.* C1-Cx revisited: intramolecular synergism in a cellulase. *Proc Natl Acad Sci USA* **91,** 11383–11387 (1994).
10. Gao, P. J., Chen, G. J., Wang, T. H., Zhang, Y. S. & Liu, J. Non-hydrolytic disruption of crystalline structure of cellulose by cellulose binding domain and linker sequence of cellobiohydrolase I from *Penicillium janthinellum*. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)* **33,** 13–18 (2001).
11. Wang, L., Zhang, Y. & Gao, P. A novel function for the cellulose binding module of cellobiohydrolase I. *Sci China C Life Sci* **51,** 620–629 (2008).
12. Giardina, T. *et al.* Both binding sites of the starch-binding domain of *Aspergillus niger* glucoamylase are essential for inducing a conformational change in amylose. *J Mol Biol* **313,** 1149–1159 (2001).
13. Cuskin, F. *et al.* How nature can exploit nonspecific catalytic and carbohydrate binding modules to create enzymatic specificity. *Proc Natl Acad Sci USA* **109,** 20889–20894 (2012).
14. Georgelis, N., Tabuchi, A., Nikolaidis, N. & Cosgrove, D. J. Structure-function analysis of the bacterial expansin EXLX1. *J Biol Chem* **286,** 16814–16823 (2011).
15. Georgelis, N., Yennawar, N. H. & Cosgrove, D. J. Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. *Proc Natl Acad Sci USA* **109,** 14830–14835 (2012).
16. Fernandez-Arrojo, L., Guazzaroni, M. E., Lopez-Cortes, N., Beloqui, A. & Ferrer, M. Metagenomic era for biocatalyst identification. *Curr Opin Biotechnol* **21,** 725–733 (2010).
17. Li, R., Kibblewwhite, R., Orts, W. J. & Lee, C. C. Molecular cloning and characterization of multidomain xylanase from manure library. *World J Microbiol Biotechnol* **25,** 2071–2078 (2009).
18. Peng, H. *et al.* A starch-binding domain identified in alpha-amylase (AmyP) represents a new family of carbohydrate-binding modules that contribute to enzymatic hydrolysis of soluble starch. *FEBS Lett* **588,** 1161–1167 (2014).
19. Duan, C. J., Liu, J. L., Wu, X., Tang, J. L. & Feng, J. X. Novel carbohydrate-binding module identified in a ruminal metagenomic endoglucanase. *Appl Environ Microbiol* **76,** 4867–4870 (2010).
20. Cheema, T. A., Jirajaroenrat, K., Sirinarumitr, T. & Rakshit, S. K. Isolation of a gene encoding a cellulolytic enzyme from swamp buffalo rumen metagenomes and its cloning and expression in *Escherichia coli*. *Anim Biotechnol* **23,** 261–277 (2012).
21. Ferrer, M. *et al.* Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environ Microbiol* **7,** 1996–2010 (2005).
22. Nguyen, N. H. *et al.* Identification and characterization of a cellulase-encoding gene from the buffalo rumen metagenomic library. *Biosci Biotechnol Biochem* **76,** 1075–1084 (2012).
23. Ferrer, M. *et al.* Functional metagenomics unveils a multifunctional glycosyl hydrolase from the family 43 catalysing the breakdown of plant polymers in the calf rumen. *Plos One* **7,** e38134 (2012).
24. Sunna, A., Gibbs, M. D. & Bergquist, P. L. Identification of novel beta-mannan- and beta-glucan-binding modules: evidence for a superfamily of carbohydrate-binding modules. *Biochem J* **356,** 791–798 (2001).
25. Abou Hachem, M. *et al.* Carbohydrate-binding modules from a thermostable *Rhodothermus marinus* xylanase: cloning, expression and binding studies. *Biochem J* **345,** 53–60 (2000).
26. Henshaw, J. L. *et al.* The family 6 carbohydrate binding module CmCBM6-2 contains two ligand-binding sites with distinct specificities. *J Biol Chem* **279,** 21552–21559 (2004).
27. Montanier, C. *et al.* Circular permutation provides an evolutionary link between two families of calcium-dependent carbohydrate binding modules. *J Biol Chem* **285,** 31742–31754 (2010).
28. Montanier, C. Y. *et al.* A novel, noncatalytic carbohydrate-binding module displays specificity for galactose-containing polysaccharides through calcium-mediated oligomerization. *J Biol Chem* **286,** 22499–22509 (2011).
29. Luis, A. S. *et al.* Understanding how noncatalytic carbohydrate binding modules can display specificity for xyloglucan. *J Biol Chem* **288,** 4799–4809 (2013).
30. Yoda, K., Toyoda, A., Mukoyama, Y., Nakamura, Y. & Minato, H. Cloning, sequencing, and expression of a *Eubacterium cellulosolvens* 5 gene encoding an endoglucanase (Cel5A) with novel carbohydrate-binding modules, and properties of Cel5A. *Appl Environ Microbiol* **71,** 5787–5793 (2005).
31. Ali, E. *et al.* Functions of family-22 carbohydrate-binding module in *Clostridium thermocellum* Xyn10C. *Biosci Biotechnol Biochem* **69,** 160–165 (2005).
32. Hamada, N. *et al.* Role of cellulose-binding domain of exocellulase I from white rot basidiomycete *Irpex lacteus*. *J Biosci Bioeng* **91,** 359–362 (2001).
33. Wang, Y., Yuan, H., Wang, J. & Yu, Z. Truncation of the cellulose binding domain improved thermal stability of endo-beta-1,4-glucanase from *Bacillus subtilis* JA18. *Bioresour Technol* **100,** 345–349 (2009).
34. Pires, V. M. *et al.* The crystal structure of the family 6 carbohydrate binding module from *Cellvibrio mixtus* endoglucanase 5a in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. *J Biol Chem* **279,** 21560–21568 (2004).
35. van Bueren, A. L., Morland, C., Gilbert, H. J. & Boraston, A. B. Family 6 carbohydrate binding modules recognize the non-reducing end of beta-1,3-linked glucans by presenting a unique ligand binding surface. *J Biol Chem* **280,** 530–537 (2005).
36. Herve, C. *et al.* Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proc Natl Acad Sci USA* **107,** 15293–15298 (2010).
37. de Vries, R. P. & Visser, J. *Aspergillus* enzymes involved in degradation of plant cell wall polysaccharides. *Microbiol Mol Biol Rev* **65,** 497–522 (2001).
38. Duan, C. J. *et al.* Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *J Appl Microbiol* **107,** 245–256 (2009).
39. Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K. & Pease, L. R. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* **77,** 51–59 (1989).
40. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227,** 680–685 (1970).
41. Andrade, M. A., Chacon, P., Merelo, J. J. & Moran, F. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein Eng* **6,** 383–390 (1993).

Acknowledgements
This work was financially supported by grants from the National Natural Science Foundation of China (grant number 31260211), the Guangxi Natural Science Foundation (grant numbers 2012GXNSFGA060005 and 2013GXNSFAA019089), the Guangxi BaGui Scholars Program Foundation (grant number 2011A001), and the Key Program of Educational Commission of Guangxi (grant number 201202ZD001).We thank Bernard Henrissat for useful suggestions on the classification of the novel family CBM.

Author Contributions
C.-J.D. and J.-X.F. contributed to the concept, design, and analysis of data, in addition to the preparation of the manuscript. Y.-L.F., Q.-L.C. and M.-Y.H. contributed to the acquisition of data. All the authors have approved the manuscript.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Duan, C.-J. *et al.* Identification of a novel family of carbohydrate-binding modules with broad ligand specificity. *Sci. Rep.* **6**, 19392; doi: 10.1038/srep19392 (2016).