**Supplementary Information** 

**Supplementary Figures** 

Supplementary Figure 1 Comparison of CBM40 crystal structures indicating conservation of the  $\beta$ -sandwich fold



(a) *Rg*NanH\_CBM40 (*Rg*CBM40) with 3'SL, (b) *Sp*CBM40\_NanA with 3'SL, (c) *Sp*CBM40\_NanB, (d) *Sp*CBM40\_NanC with Neu5Ac, (e) *Md*CBM40\_NanL, (f) *Cp*CBM40\_NanI with 3'SL, (g) *Cp*CBM40\_NanJ with Neu5Ac, and (h) *Vc*CBM40\_NanH with Neu5Ac. Ligands are represented by black sticks. For the *Cp*CBM40\_NanI image the loop indicated to form additional water mediated hydrogen bonding interactions with the sialyllactose galactose residue is indicated with a star. Supplementary Figure 2 Comparison of CBM40 sialic acid binding sites



(a) RgCBM40 with 3'SL, (b) SpCBM40\_NanA with 3'SL, (c) SpCBM40\_NanB, (d)
SpCBM40\_NanC with Neu5Ac, (e) MdCBM40\_NanL, (f) CpCBM40\_NanI with 3'SL, (g)
CpCBM40\_NanJ with Neu5Ac, and (h) VcCBM40\_NanH with Neu5Ac. Ligands are
represented as black sticks with hydrogen bonding interactions as black dashed lines.

**Supplementary Figure 3** Comparison of the Neu5Ac glycerol environment of the ligand in complex with CBM40 structures



(a) *Rg*CBM40 (b) *Sp*CBM40\_NanA, and (c) *Cp*CBM40\_NanI. Ligands and interacting residues are coloured. Dashed black lines indicate hydrogen bonding interactions.

**Supplementary Figure 4**. Alignment of a subset of the domain hits of the combined-CBM40 pHMM



This alignment consists of 51 nonredundant domain sequences from the complete alignment of all domain hits produced by hmmsearch of HMMER3, using the combined-model (canonical and Vibrio-type CBM40) as the query. Refer to **Supplementary Methods** for a detailed description of how this set of 51 was obtained. These were used to create the tree in **Fig 3.** Sequence identifiers are abbreviated versions of those in **Fig. 3**. The top 43 sequences are putative canonical CBM40 and the bottom 8 *Vibrio*-type. Note the position of the *Actinobacillus* (Gammaproteobacteria; canonical CBM40) sequence, 9<sup>th</sup> from bottom. Positions of the conserved binding sites in this alignment are as follows (with corresponding positions of the alignment in **Fig. 2** shown in parentheses). The site position common to both canonical and *Vibrio* types is 96 (119 in **Fig. 2**). Canonical-only sites: 78 (104), 102 (125), 104 (127), 116 (135), 118 (137), 228 (226), 241 (233). *Vibrio* type-only sites: 39 (66), 41 (68), 79 (105), 218 (216), 235 (231), 240 (232).

**Supplementary Figure 5** Analysis of STD NMR spectra of the binding of (**a**) Neu5Acα2-3Lac (3'SL) and (**b**) Neu5Acα2-6Lac (6'SL) to *Rg*CBM40



The figure shows <sup>1</sup>H,<sup>13</sup>C HSQC spectra of both ligands and the STD NMR spectra (top, red thick lines) superimposed to the 1D <sup>1</sup>H NMR reference spectrum (top, blue thin lines). To facilitate the comparison, the reference spectra are shown at 1/16 of their intensity. HSQC data were fundamental to elucidate the identity of the ligand protons receiving saturation. In both cases, the STD NMR spectra showed that the most intense STD signals corresponded to the protons of the non-reducing terminal sialic acid rings (signals assigned, red dot-dashed line in the HSQC spectra).

**Supplementary Figure 6** Binding epitope mapping of sialoglycans bound to *Rg*CBM40 as determined by STD NMR



(a) Neu5Gc $\alpha$ 3Lac (3'SLGc), (b) Neu5Gc $\alpha$ 6Lac (6'SLGc), (c) Neu5Ac $\alpha$ 3Gal (3'SGal), (d) Neu5Ac $\alpha$ 6Gal (6'SGal), (e) Neu5Ac $\alpha$ 3LacNAc (3'SLN), (f) Neu5Ac $\alpha$ 6LacNAc (6'SLN), (g) Neu5Ac $\alpha$ 6Gal $\alpha$ OC<sub>3</sub>H<sub>6</sub>N<sub>3</sub> (Neu5Ac-STn), (h) Neu5Gc $\alpha$ 6Gal $\alpha$ OC<sub>3</sub>H<sub>6</sub>N<sub>3</sub> (Neu5Gc-STn), and (i) Neu5Ac $\alpha$ 3Gal $\beta$ 3GalNAc $\alpha$ OC<sub>3</sub>H<sub>6</sub>N<sub>3</sub> (STF). Legend indicates relative STD intensities normalised at H7: blue, 0–24%; yellow, 25–50%, red 51–100%; larger red dots indicate values over 100%. STD NMR spectra of the binding of 3'SLGc and Neu5Gc $\alpha$ 2-6Lac 6'SLGc to *Rg*CBM40 are shown as a representative example.



Supplementary Figure 7 CD spectra of RgCBM40 wt and mutants

(**a**) WT (light green), I95A (dark purple), Y116A (red), E126A (dark green), R128A (dark blue), Y210A (light blue), boiled WT (light purple). (**b**) WT (dark blue), boiled WT (green), R204A/R128A (brown), R204A-light blue.

## Supplementary Figure 8 ITC isotherms of *Rg*CBM40 to sialoglycans



(a) *Rg*CBM40 R128A/R204A binding to 3'SL, (b) *Rg*CBM40 I95A binding to 3'SL. The Kd is indicated in mM. NB indicates no binding detected. Traces shown are representative examples.

**Supplementary Figure 9** Binding epitope mapping of sialoglycans bound to *Rg*CBM40 I95A as determined by STD NMR



(a) 3'SL and (b) 6'SL. Legend indicates relative STD intensities normalised at H7: blue, 0–24%; yellow, 25–50%, red 51–100%; larger red dots indicate values over 100%.

**Supplementary Figure 10** Substrate specificity of *Rg*NanH and *Rg*GH33 as analysed by HPAEC-PAD



The substrate ( $\alpha$ 2-3 sialyllactose-3'SL,  $\alpha$ 2-3 Lewis X-3S'LX or LS174T MUC2-MUC2) was incubated with *Rg*NanH or *Rg*GH33 and the reaction products analysed by HPAEC-PAD. For 3'SL and 3'SLX the % of 3'SL and 3'SLX remaining respectively is plotted. For MUC2, the % of Neu5Ac remaining attached to the MUC2 after removal of the sugars which have been enzymatically liberated is plotted.

**Supplementary Figure 11** Immunodetection of IT-sialidase on the cell surface of *R. gnavus* strains



(a) Immunogold labelling of *R. gnavus* strains ATCC 29149 and E1. *R. gnavus* strains ATCC 29149 and E1 grown on 3'SL or glucose, respectively, were probed with anti-*Rg*NanH antibody and analysed by transmission electron microscopy (TEM). In each image the scale bar represents 100 nm. (b) Western blotting of *R. gnavus* expressed proteins. *R. gnavus* ATCC 29149 and E1 were grown on 3'SL or glucose, respectively. Proteins isolated from the cell wall (CW), cytoplasm (CP), and extracellular proteins from the spent media (SM) were analysed by SDS-PAGE (left) and western blot using anti-*Rg*NanH antibody (right). (M) – Broad Range, Blue Protein Standard (NEB).

**Supplementary Figure 12** ELISA of *Rg*CBM40 at different concentrations binding to a range of purified mucins



Legend refers to mucin 2 (MUC2) and mixed mucins (mucins) from human cell line LS174T, purified pig gastric mucin (pPGM), mucin 2 (muc2) and mixed mucins from small intestine (SI) and colon (C) of wild type (WT), C3GnT<sup>-/-</sup> and germ free (GF) mice. The error bars show the standard error of the mean (SEM) of three replicates.

**Supplementary Figure 13** A wall-eye stereo image of a portion of the 2Fo-Fc electron density map of the *Rg*CBM40 3'SL complex X-ray crystal structure



The map is contoured to 2  $\sigma$  (0.89e/A3) and is centred on Phe72.

#### Supplementary Figure 14 CBM40 domain sequence alignment used to create the pHMMs

		10	20	30	40	50	éo	70	
SpCBM40_NanA		VE T	VI <mark>E</mark> KEDVE	T - N A S N <mark>G</mark> Q					RV
A0A0U0KEN0_STREE SoCBM40_NanB		T	LEOGGSYO	NN K S					RV
S2XRE8_9STAP		<u>.</u>		LKG-G-G-					j
SpCBM40_NanC		<mark>P</mark>							E
CpCBM40 Nani	LNVYE	IKGEVD E	LANYGNLK	TKEEER-					V
L10888_9CLOT	VNIYR	VKEDIN K	IARYENKK	1 TN - NQ G -					· <mark> </mark>
MdCBM40_NanL		PEGI		I - AEGQG -					· · · · · · Č
RgCBM40		SVP	VLQKEGIE	I-SEGTG-					<mark></mark>
E5XHL0_9FIRM			VF RKEDQK	I-AVGSP-					V
VcCBM40_NanH A0A084TBH4 9VIBR		S-NAA	LFDYN	ATTD PVKD	SPARQGWMQD1	HSOSGSGA	L I N A	GVTAWVA0GEG6	RAQWIY
A0A162HC21_9VIBR		A - N A E	L F <mark>D</mark> Y T	A T T D <mark>P</mark> A K D	Q P S E Q G W I A D H	⊣sǫs <mark>s</mark> ssv	LSTVC	GVTAWVAQGEGG	RAQWKA
DOZ2P3_PHODD		AH	VYQFI	PSDNPTND			AVVN 0		RANWTI
A0A135/7 Y9_9GAMM		SAE		SENPQFD	E PIEHGWDNYI	LSGSAYGE	AVKK	KGVEWFVNGED	RANWKI
A8H2Q3_SHEPA			A	TFDSNLNT	PLLAGWTND	SVDSGAGS	LVNE 0	GQAAWQARGING	RAEWEI
A8H2Q3_SHEPA A0A0O0H4O4 9GAMM		A-NIL	• A		SPDLLGWTND	STLOGOGS	VAQD LLDD N	GDTVW0ADGSAG	KAEWEV
A0A081KGM9_9GAMM				- FNAESDS	S <mark>P</mark> IAQ <mark>G</mark> WS <mark>G</mark> NI	LSGPAYGE	LVEETLAD <mark>G</mark> T	TVLAWYVNGLEG	IAEWKA
		90	100	110	120	130	140	150	
SpCBM40_NanA	D L S S - E L	DKLKKLENA	<mark>Т </mark> ИНМЕ	DAK <mark>AP</mark> AFYI	NLFS <mark>V</mark> SSA - TI	K KD E YF	TMAVYNN-TA	T L EG R G S D G K Q F	YNNYND
A0A0U0KEN0_STREE	DL S-DEL	EKVQGLQNA		AADGPSFY	NLFSASST-TH				YGSYTD
S2XRE8_9STAP	NVTEDIL	DKL - NGPGF	TVIIKYSQ	SKP-DGVQ	ALFGISNS-K	KGNANSYL	DLYINEKGEL	GMEARD - SGSQT	
SpCBM40_NanC	NVTKELK	DKFTS <mark>GD</mark> -F	<mark>τννικγ</mark> ΝQ	S <mark>S</mark> E - K <mark>G</mark> LQ	ALF <mark>GISN</mark> S-K <mark>I</mark>	PGQQNSYV	D V F L R D N <mark>G</mark> E L	. GMEARD TSSN	KNNLVS
AUA166NLZ1_STAPS CpCBM40 Nani	NITSGAL	NKL - IGPDF	TIVTRENM	- ND - TS IO	SLIGLSDG - NI	KUNPSSYL KANNYE		GYEL RROEGNOD	
L1Q888_9CLOT	DI SS - DI	EALKGLEEG	TIVARFDT	TK <mark>G</mark> DIQ	SLIS <mark>VGN</mark> N-N	V AN <mark>G</mark> HF	HLYVAD - NT	GFEVRNQSGN -	- VATGK
MdCBM40_NanL	SLDQEA-GA	KYVKAMTQG	TIILSYKS	TSE-NGIQ		AGNODRHF	HIYITNSGGI		FNYTLD
RgCBM40	DLSKEP-GA	ATVKALEQG	TIVISYKT	TSE-NAIQ	SLLSVGNG - TH	KGNQDRHF	HLYITNAGG	GMEL RNTDGE	FKYTLD
E5XHL0_9FIRM	ELS-TKDYA	EKLMKLEEG	TILVHYTS	TSD-QAIQ	SLFSVSNA-KA	AGHENRHF	HVYIR <mark>P</mark> EGVL	GCEIRNESA	MN Y <mark>G F</mark> K
VcCBM40_NanH A0A084TBH4 9VIBR	VL - SDALK	SOATN - YGW	RMTTEMKV BLSTEMBV	L 5G GM V 5G GM	TNYYANG	TQRVLP SORVLP	I I S L D S S G N L		
A0A162HC21_9VIBR	VL PDALK	SQAAA - Y <mark>GW</mark>	RLSTEMRV	∨ <mark>sg</mark> gM	ITNYYADG	SQRVLP	ILSINSAGEL	VVEF EGRSG	T TVL
DOZ2P3_PHODD	TP TTELN	NQATR - YGW		V <mark>SG</mark> GY	I TNYYANG	QHRFLP			
A0A135/7 Y9_9GAMM	TP TDALN	DQATA - FGW	TMSAEMRV	KKGGY	LTNYYANG	EYRFLP	IISLQ-DNQL	TAEF EGEDQ	
A8H2Q3_SHEPA	Q P SVQVN	RD <mark>A</mark> TT - L <mark>GW</mark>	KMTTTS <mark>R</mark> I	V <mark>S G</mark> S A	I SDY <mark>Y</mark> A <mark>NG</mark>	NKRFLA	LLSINGNGDL	VAAL EGGSS	3H <mark>T L</mark> V
A8H2Q3_SHEPA A0A0O0H4O4 9GAMM	VP TELN	INEASL - KGW	RMSWTSRV SMSSTVKV	ESGGY ISGSY	TDYYANG	ALRFL <mark>P</mark> NKRYLV	ILSISASGDL NIKIDSSGAL	VVNL VGGGE	
A0A081KGM9_9GAMM	TPEGEFN	AMASQ - NGW	TMAFTMRV	E <mark>S G</mark> S Y	LTNY <mark>YGNG</mark>	V K R F <mark>L P</mark>	VLKLD - GNNL	VVEL EGDQA	YTLV
		170	180	190	200	210	220	230	
SpCBM40_NanA	A <mark>P</mark> L - K <mark>V</mark> K <mark>P</mark> -	<mark>G</mark> QW <mark>N</mark> S	V <sup>†</sup> F⊺V <mark>E</mark> K <mark>P</mark>	TAEL <mark>P</mark> K <mark>G</mark> R	VRL <mark>YV</mark> NGVLS	R T - S L	R S <mark>G</mark> NFIK DM <mark>F</mark>	<sup>•</sup> D∨TH <mark>V</mark> Q <mark>IG</mark> ATKF	RAN - NTV
A0A0U0KEN0_STREE	APL-KIRP-	GKYNS	VTFTVERP	RKDS <mark>P</mark> NGQ		RTN K	KSGKFLADM	DVDKLQLGATNF	A-GELK
S2XRE8_9STAP	RPA-SVWGK	YKNKPVSNT	VALVSNNN	SN T	YALYSNGYKI	EEKKL	DKFLKLNDIK	GLDSIVIGGVNF	E-GTNK
SpCBM40_NanC	R <mark>P</mark> A - S <mark>V</mark> WGK	YKQEAVTNT	VAVVADSV	ккт	Y <mark>SLYANG</mark> TKV	VEKK V	DN <mark>F</mark> LNIKDIK	( <mark>GIDY<mark>Y</mark>MLGGVKF</mark>	RA-GK <mark>T</mark> A
A0A166NLZ1_STAPS CoCBM40 Nani	ADV-TF	-NRGINT	VALVLNHY	TKT	AKIFING YKVE	EEKK I KTVS D	NNFLMLSNIN PNIKFINAI -	NINSGELGKTDE	RE-GKNS
L1Q888_9CLOT	ASA-VL	- NNG I NT	VALKVTS-	<mark>G</mark> V G	KIFINGKLA	GEVTT	SNATLSAGV	DVNNAYIGKTDF	AS <mark>G</mark> -NE
MdCBM40_NanL	RPA-SVRAL	YKGERVENT		NKQ	CRLFANGELLA	ATLD K	DAFKFISDI	GVDNVTLGGTKF	
RgCBM40	CPA-AVRGS	YKGERVSNT	VALKADKE	N <u>K</u> Q	KLFANGELI	ATLD Q	EAFKFISDI	GVDNVMLGGTMF	Q - GTVA
E5XHL0_9FIRM	AAN - AVKAD	YKGKPAENI	IALQADKE	к <mark>с</mark> т	YQLFA <mark>NG</mark> EKVI	LTVDAAAL	GGYRFISEI	<mark>G</mark> LDT <mark>VSLG</mark> ATKF	
A0A084TBH4 9VIBR	ATG-AAATD	YHK	YELVFLPG	5NP DNPN	ASFYFDGKLII	LD01S	PSP	SONMLVWGN	GST
A0A162HC21_9VIBR	ATG-AAATD	<mark>Ү</mark> нк	YEL∨FH <mark>PG</mark> I	D А <mark>Р</mark> К	A <mark>S</mark> F <mark>Y</mark> F D <mark>G</mark> Q L I I	LDQ   S	P S P 1	SQ <mark>NMLVWG</mark> N	<mark>G S</mark> Т
DUZ2P3_PHODD A0A0B7IGA0_PHOPO	ASG-DDVHG AAD-STVND	Yнк Yнг	YDIVEHPG	<mark>Р N Р</mark> Т. D Т O Т	ASFFFDGVLII	K SDW Q RDOW V	G H P	SQNMIAWGN	GSS
A0A135/7 Y9_9GAMM	ATG-RDAKK	YHQ	YD I VFH PG	Е Т <mark>Р</mark> Т	ASFYFDGELI	KS <mark>G</mark> W E	P T P 1	TQ <mark>N</mark> MVVW <mark>G</mark> N	<mark>G S</mark> S
A8H2Q3_SHEPA	PMTGADK	YHT	YEVTYDVE	T - S L	ATFYFDGIAIE	E - TW A	G S A 1	SQNVIVWGN	<mark>GS</mark> S
A0A0Q0H4Q4 9GAMM	SQ QG SDQ	YHQ	YEVNYDAS	S-QQ	ATFWFDGEKV	T - SW S	G S A	NONVIVEGN	GSS
A0A081KGM9_9GAMM	EGDLEAAMG	YHN	Y V I S F <mark>D P</mark> A	Т - <mark>G</mark> К	ATYQFDDKII	TSDW Q	G S E 5	AHNVVAWGN	<mark>G S</mark> S
		250	260	270					
SpCBM40_NanA	W-GSNLQ	IRNLTVYNR	ALT <mark>P</mark> EEVQ	KRSQLFK					
AUAOUOKENO_STREE SpCBM40 NanB	W-GSDLS Υ-LAκας	IDELS FNK	ALTPEEV-	TIPLSNP					
S2XRE8_9STAP	F-GFNGT	IENIKIYNY	PLNEKNLE	NKTQNN -					
SpCBM40_NanC	F-GFNGT	LENIKFFNS	ALDEETVK	KMTTNA -					
CpCBM40_Nanj	Y - L F RGN	I D F MN I YD K	PVSDNYLL	RKTGETK					
L1Q888_9CLOT	Y - P F S G S	IDFIDVYGD	VL <mark>P</mark> DKYLL	DVT <mark>G</mark> QT -					
MdCBM40_NanL C9L8Y9 BL4H4	Y-PFGGT		ALSDEELI PLSDEELI	QATGVTT ESTGKTT					
RgCBM40	Y - PF GGS	IERMQVYRD	VLSDDELI	AVTGKT -					
E5XHL0_9FIRM	Y-TFGGN	THKIEVYET	PLTDEELI	ЕЕТККТ -					
A0A084TBH4_9VIBR	TDG VSA	YKNIEFEVQ	GD						
A0A162HC21_9VIBR	TDG VSA	YKNIEFEVQ	GD						
DUZ2P3_PHODD A0A0B7IGA0_PHOPO	SIAG EAY	YKSIAFOVN	GD						
A0A135/7 Y9_9GAMM	HVDA - KAY	YRSVNFTIS	G E						
A8H2Q3_SHEPA	GTNG VAN	YRYVN FETF	G						
A0A0Q0H4Q4_9GAMM	GTSG - VAN	YKNVRFEV-							
A0A081KGM9 9GAMM	STSGASEAF	YQSFTFDIP	G						

This is based on the structural alignment (see text and **Fig. 2**), with additional homologous segments selected from the result of Blast searches of the Uniprot database. These hits (one for each of the canonical and 9 for the *Vibrio*-type) were selected on the basis of balancing lowest sequence identities possible and a range of taxa while maintaining most or all of the conserved CBM40 features (see **Fig. 2**). Most of the proteins are annotated as 'sialidase'. The 'combined', canonical and *Vibrio*-type pHMMs were constructed using HMMER3 from respectively all, the first 12, and the last 10 sequences. Refer to the text for the key to the 7 original sequences. The remaining 14 sequence names are UniProt identifiers.

### **Supplementary Tables**

<i>Rg</i> CBM40	Neu5Ac	3'SL	6'SL	Neu5Gc	3'SLGc	6'SLGc
WT	21*	0.57 ± 0.05	1.70 ± 0.14	21*	2.69 ± 0.86	11*
195A	23*	1.82 ± 0.13	1.37 ± 0.18	NT	NT	NT
Y116A	NB	NB	NB	NT	NT	NT
E126A	NB	NB	NB	NT	NT	NT
R128A	NB	NB	NB	NT	NT	NT
R204A	NB	NB	NB	NT	NT	NT
R128A/R204A	NB	NB	NB	NT	NT	NT
Y210A	NB	NB	NB	NT	NT	NT

**Supplementary Table** 1 Kd (mM) of *Rg*CBM40 wild type (WT) and mutants for different sugars, as determined by ITC

The cell contains 115-230 µM protein and the syringe 10 mM sugar (25 mM for Neu5Ac).

The error reported is the standard deviation of three results.

\* These values are estimates only as the Kd is too high to determine with the concentration of sugar used.

NB- no binding detected, Kd >>25 mM NT-not tested

**Supplementary Table 2.** Thermodynamic parameters for *Rg*CBM40 WT and I95A binding to 3'SL and 6'SL

<i>Rg</i> CBM40 + sugar	ΔH (kcal/mol)	ΔG (kcal/mol)	-T∆S (kcal/mol)
WT + 3'SL	-5.90 ± 0.72	$-4.43 \pm 0.05$	1.47 ± 0.77
195A +3'SL	-4.71 ± 0.43	$-3.74 \pm 0.04$	$0.97 \pm 0.47$
WT + 6'SL	-9.61 ± 0.44	$-3.78 \pm 0.05$	$5.82 \pm 0.48$
195A + 6'SL	$-3.79 \pm 0.24$	-3.91 ± 0.07	$-0.12 \pm 0.3$

Enthalpy ( $\Delta$ H), Gibbs free energy ( $\Delta$ G) and entropy (-T $\Delta$ S) values are shown. The error reported is the standard deviation of three results.

91
70
59
62
34
24
8
2
0

Supplementary Table 3. Sialic acid levels in purified mucins as determined by MS

The mucins are mucin 2 (MUC2) and mixed mucins (mucins) from human cell line LS174T, purified pig gastric mucin (pPGM), and muc2 from germ free (GF), wild type (WT), and *C3GnT*<sup>-/-</sup> mutant mice.

Protein encoded	Direction	Sequence (5'-3')
<i>Rg</i> NanH WT	F	GATATCGGATCCCAAGAGGCCCAGACAG
	R	TGGTGCTCGAGTTATGGTTGAACTTTCAGTTCATC
<i>Rg</i> CBM40 WT	F	GATATCGGATCCGTGTTGCAAAAGGAAGGAATC
	R	GGTGCTCGAGTTACTTTCCTGTCACAGCAATAAG
<i>Rg</i> GH33 WT	F	GATATCGGATCCAATATCTTTTATGCAGGAGATGC
	R	TGGTGCTCGAGTTATGGTTGAACTTTCAGTTCATC
<i>Rg</i> GH33 D282A	F	GTATTGATCTTACTTTTTGCAGCATGGGTGCCACCATATCTTG
	R	CAAGATATGGTGGCACCCATGCTGCAAAAAGTAAGATCAATAC
<i>Rg</i> CBM40 195A	F	CAACCAGTGAAAATGCGGCTCAATCGTTATTGAGTG
	R	CACTCAATAACGATTGAGCCGCATTTTCACTGGTTG
<i>Rg</i> CBM40 Y116A	F	GATAGACATTTCCACTTAGCTATCACAAATGCAGGCGG
	R	CCGCCTGCATTTGTGATAGCTAAGTGGAAATGTCTATC
<i>Rg</i> CBM40 E126A	F	GCGGCGTAGGTATGGCATTGAGAAATACAG
	R	CTGTATTTCTCAATGCCATACCTACGCCGC
<i>Rg</i> CBM40 R128A	F	GCGTAGGTATGGAATTGGCAAATACAGATGGCGAG
	R	CTCGCCATCTGTATTTGCCAATTCCATACCTACGC
<i>Rg</i> CBM40 R204A	F	GTAATGCTGGGCGGTACCATGGCTCAGGGAACCGTTGCCTATC
	R	GATAGGCAACGGTTCCCTGAGCCATGGTACCGCCCAGCATTAC
<i>Rg</i> CBM40	F (R128A)	GCGTAGGTATGGAATTGGCAAATACAGATGGCGAG
R128A/R204A	F (R204A)	GTAATGCTGGGCGGTACCATGGCTCAGGGAACCGTTGCCTATC
	R (R128A)	CTCGCCATCTGTATTTGCCAATTCCATACCTACGC
	R (R204A)	GATAGGCAACGGTTCCCTGAGCCATGGTACCGCCCAGCATTAC
<i>Rg</i> CBM40 Y210A	F	CAGGGAACCGTTGCCGCTCCATTTGGAGGTTC
	R	GAACCTCCAAATGGAGCGGCAACGGTTCCCTG

Supplementary Table 4. Primer sequences used in this study

The direction of the primers is either forward (F) or reverse (R).

#### **Supplementary Methods**

Construction of the domain sequence profile models representing different types of CBM40 domains, and comparison with existing models. Models (pHMMs) are depicted as truncated boxes. Numbers of hits, including the Venn diagrams, are numbers of domains (not protein sequences) which match each model with an i-Evalue of  $\leq 10^{-6}$  when searching the 176,818,559 protein sequences obtained from 67,248 prokaryote genomes (NCBI, April 2016), using the HMMER3 software. (a) Alignment produced from a manual inspection of the 3D structures of seven CBM40 domains, including VcCBM40 (see main manuscript text), supplemented with a selection of homologous domain sequences from the UniProt database, giving an alignment of 12 canonical and 10 Vibrio-type CBM40s (see Supplementary Fig. 14). This was also divided into two components (canonical-only and Vibrio type-only), and all three were used to produce profile Hidden Markov Models using HMMER3. (b) The principal Pfam family matching known canonical CBM40s is "Sialidase, N-terminal domain" (abbreviated in Pfam to "Sialidase" which is potentially confusing, and has therefore been referred here as "Sialidase(NTD)". Furthermore, around half of the Sialidase(NTD)-matching domains also matched a second Pfam family, "Laminin G 3" (also known as "Concanavalin A-like lectin/glucanases") - which is a member of the same wider superfamily (clan) as Sialidase(NTD). This should not be confused with proteins having two domains in different parts of the sequence, each matching a different family – which does occur in many cases. (c) Our own canonical CBM40 model performs near identically to Sialidase(NTD), while finding a few additional matches. We consider the approximately 16,000 domains matching either canonical CBM40 or Sialidase(NTD) as the 'canonical CBM40 hits', and the approximately 13,000 which match only Laminin G 3 as 'LG3-only'. (d) Unlike the canonical CBM40s, there appeared to be previously no domain model which strongly matches the Vibrio-type CBM40. However, some domain hits to our new Vibrio-type model also match the Pfam family "Siallect-inser" (also known as "Vibrio cholerae sialidase, lectin insertion"). Our Vibrio-type model captures almost all hits of Sial-lect-inser, as well as many others not matched by it. Analogous to the CBM40/Laminin\_G\_3 incidence, a number of proteins contained two domains, one matching each of the Vibrio-type CBM40 model, and the Sial-lect-inser. (e) The new combined CBM40 model succeeds (red ovals) in matching most Vibrio type and Sial-lect-inser domains as well most canonical CBM40s, while excluding almost all of the LG3-only domains.



Full breakdown of the hits of the combined-model. The numbers of domain hits for each pHMM in the database of 176,818,559 protein sequences obtained from 67,248 annotated prokaryote genomes, was determined using HMMER3 (hmmsearch) with a threshold of i-Evalue of  $\leq 10^{-6}$ . See above for details of each model. The numbers of hits which match 1 or more of the models are shown in a single Venn diagram. This illustrates that at this significance cut-off, there was no domain matching both a canonical and Vibrio type. However, some combined-model hits (7,146 + 18) matched both the Sialidase(NTD) and Laminin G 3, which both belong to the same superfamily (see above). Similarly, 331 combined-model hits matched both the Vibrio-type CBM40 and Sial-lect-inser. The 16,657 combined-model hits consist of only 396 unique sequences, and these were classified based on the four sub-types (canonical CBM40, Laminin\_G\_3, Vibrio-type CBM40, Sial-lect-inser), with a manual evaluation of binding sites resolving E-value "ties" or near-ties in a few (18) cases. Discarding the non-CBM40 types resulted in 352 CBM40 hit sequences (canonical and Vibrio type), of which 3 were excluded (as representing partial matches to the query pHMM). Application of a redundancy threshold of  $\leq$  80% sequence identity (using Jalview<sup>1</sup>) resulted in 51 sequences (Supplementary Fig. 4) used for the phylogenetic analysis (Fig. 3).



# Supplementary reference:

1. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191 (2009).