Supporting Information for:

Identification of glycoproteins isolated from extracellular polymeric substances of full-scale anammox granular sludge

Marissa Boleij[1], Martin Pabst[1], Thomas R. Neu[2], Mark C.M. van Loosdrecht[1], Yuemei Lin[1*]

[1] Department of Biotechnology, Delft University of Technology, van der Maasweg 9, 2629 HZ, Delft, The Netherlands

[2] Department of River Ecology, Helmholtz Centre for Environmental Research – UFZ, Brueckstrasse 3A, 39114 Magdeburg, Germany

This Supporting Information contains: 12 pages with 3 supplemental materials and methods and 8 supplemental figures. In addition, 4 supplemental tables provided separately in excel files.

[*] **Corresponding author:** Yuemei Lin, Yuemei.Lin@tudelft.nl

Content Supplemental material

Supplemental materials and methods

Supplemental Figures

Supplemental Tables (Supplementary Tables S1-4 are provided in one separate excel file)

Table S1. Bacteria wide database analysis

Table S2. Human and Crapome database analysis

Table S3. Brocadia database analysis

Table S4. WP_070066018.1 focused analysis

Supplemental materials and methods

1. Clone library analysis

DNA was extracted from sludge sample with FastDNATM 2 mL SPIN Kit for Soil (MP Biomedicals, USA). 16s rRNA gene was amplified from total DNA (forward primer 5'-RCATGCAAGTCGAACGWG-3' and reverse primer 1492R: – 5'-ACCTTGTTACGACTT-3') with Taq DNA polymerase using the following PCR condition: 5 min initial denaturation at 94°C, 25 cycles of 1 min at 94 °C, primer annealing for 1 min at 50°C and 70s at 72°C. Final extension was carried out for 4 min at 72°C. The PCR products were checked on a 1% electrophoresis agarose gel and cloned by using the TOPO TA cloning system. For each sample, 30 colonies were picked and purified for Sanger sequencing. Full length sequences were compared with similar sequences of references species by BLASTN search and grouped into same OTU based on highest percentage similarity. The sequences were checked for chimeric artefacts by decipher online tool (http://decipher.cee.wisc.edu/FindChimeras.html). Sequences were aligned by SINA online tool (https://www.arb-silva.de/).

2. Fluorescent In Situ Hybridization (FISH)

Fluorescent In Situ Hybridization (FISH) was performed as described by Johnson et al (2009). Before fixation, the granules were homogenized using a potter homogenizer. The following FISH probes were applied: EUB338 (Amann et al., 1990) to target all bacteria, Amx820 (Schmid et al., 2001) to target "*Candidatus* (*Ca*.) Brocadia" and "*Ca.* Kuenenia" species, Kst157 (Schmid et al., 2001) to target "Ca. Kuenenia stuttgartiensis" and Bfu613 (Kartal et al., 2008) to target "Ca. Brocadia fulgida". The TestProbe tool from SILVA (https://www.arb-silva.de/search/testprobe/) confirmed that Bfu613 can also target with "*Ca*. Brocadia sp.40", that is renamed to "*Ca*. Brocadia Sapporoensis". Details on oligonucleotide probes are available at probeBase (Greuter et al., 2016).

References:

Amann, R. I., Binder, B. J., Olson, R. J., Chisholm, S. W., Devereux, R., and Stahl, D. A. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl. Envir. Microbiol.* 56, 1919–1925. Available at: http://aem.asm.org/cgi/content/long/56/6/1919.

Greuter, D., Loy, A., Horn, M., and Rattei, T. (2016). probeBase—an online resource for rRNA-targeted oligonucleotide probes and primers: new features 2016. *Nucleic Acids Res.* 44, D586–D589. doi:10.1093/nar/gkv1232.

Johnson, K., Jiang, Y., Kleerebezem, R., Muyzer, G., and van Loosdrecht, M. C. M. (2009). Enrichment of a mixed bacterial culture with a high polyhydroxyalkanoate storage capacity. *Biomacromolecules* 10, 670–6. doi:10.1021/bm8013796.

Kartal, B., Van Niftrik, L., Rattray, J., Van De Vossenberg, J. L. C. M., Schmid, M. C., Sinninghe Damsté, J., et al. (2008). Candidatus "Brocadia fulgida": An autofluorescent anaerobic ammonium oxidizing bacterium. *FEMS Microbiol. Ecol.* 63, 46–55. doi:10.1111/j.1574-6941.2007.00408.x.

Schmid, M., Schmitz-Esser, S., Jetten, M., and Wagner, M. (2001). 16S-23S rDNA intergenic spacer and 23S rDNA of anaerobic ammonium-oxidizing bacteria: Implications for phylogeny and in situ detection. *Environ. Microbiol.* 3, 450–459. doi:10.1046/j.1462-2920.2001.00211.x.

3. Parameters database searches

PEAKS Studio 8.5 analysis using a bacteria-wide sequence database (Bacteria proteome collection, downloaded from UniprotKB, containing 61994821 Protein sequences, unreviewed and reviewed).

Database search was performed using the following parameters: Parent Mass Error Tolerance: 10 ppm, Fragment Mass Error Tolerance: 0.5 Da, Precursor Mass Search Type: monoisotopic, Enzyme: Trypsin, Max Missed Cleavages: 1, Non-specific Cleavage: none, Fixed Modifications: Carbamido-methylation, Variable Modifications: Oxidation, Max Variable PTM Per Peptide: 1, FDR Estimation: Enabled, Precursor Options: corrected, Charge Options: no correction. Peptide -10lgP threshold was set to ≥42.2, Peptide Ascore ≥0, Protein -10lgP ≥40, De novo ALC Score ≥80%. This resulted in identifications with a FDR (Peptide-Spectrum Matches) 0.0%, FDR (Peptide Sequences) 0.0%, FDR (Protein) 1.9%.

PEAKS Studio 8.5 analysis using a Candidatus Brocadia protein focused database (containing 16631 Protein sequences and the retracted WP_070066018.1, unreviewed and reviewed).

Database search was performed using the following parameters: Parent Mass Error Tolerance: 10 ppm, Fragment Mass Error Tolerance: 0.5 Da, Precursor Mass Search Type: monoisotopic, Enzyme: Trypsin, Max Missed Cleavages: 1, Non-specific Cleavage: one, Fixed Modifications: Carbamido-methylation, Variable Modifications: Oxidation and Deamidation, Max Variable PTM Per Peptide: 1, FDR Estimation: Enabled, Precursor Options: corrected, Charge Options: no correction. Peptide -10lgP threshold was set to ≥37.1, Peptide Ascore ≥0, Protein -10lgP ≥40 with 1 unique peptide min, De novo ALC Score ≥80%. This resulted in identifications with a FDR (Peptide-Spectrum Matches) 0.0%, FDR (Peptide Sequences) 0.0%, FDR (Protein) 0.0%.

X!tandem (release ALANINE, The GPM) was used as confirmatory database search engine using the following parameters: Reversed sequences yes, Check parent ions for charges 1, 2 & 3 yes, Protein entries with a peptide log(e) < -3 and protein log(e) < -5, Fragment mass error of 0.5 and +/-10 ppm parent ion mass error (LTQ-Orbitrap), isotope error yes, Fragment type, monoisotopic, Parent mass detection no, max parent charge: 4, minimum parent M+H= 500 Da, Minimum fragment ion = 150 Da, Total peaks: 50, minimum peaks 15, Fragment types: b/y ions, Fixed modification: Carbamidomethyl, Variable modification none, refinement no, Cleavage site Trypsin, Missed cleavage sites allowed 1, Semi-cleavage was used optional, Glycan mass was included optionally into database searches.
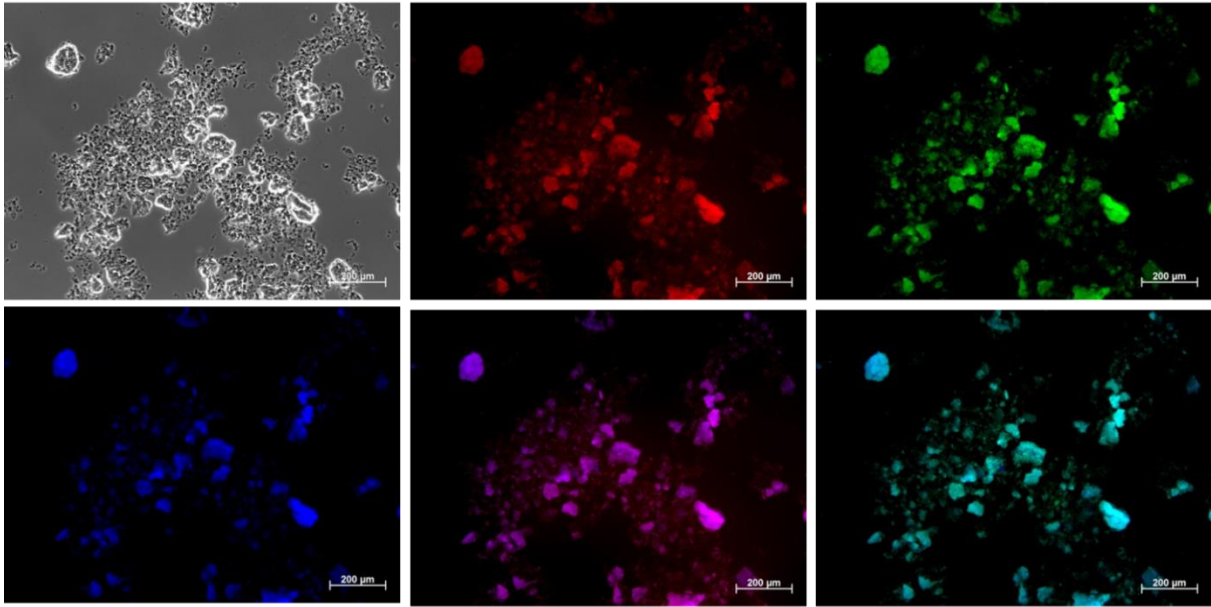
Supplemental figures

Figure S1. Fluorescent In Situ Hybridization (FISH) was performed as described by Johnson et al (2009). Before fixation, the granules were homogenized using a potter homogenizer. The following FISH probes were applied: EUB338 (Amann et al., 1990) to target all bacteria, Amx820 (Schmid et al., 2001) to target "*Candidatus* (*Ca*.) Brocadia" and "*Ca.* Kuenenia" species, Bfu613 (Kartal et al., 2008) to target "*Ca*. Brocadia fulgida" and Kst157 (Schmid et al., 2001) to target "*Ca*. Kuenenia stuttgartiensis". Details on oligonucleotide probes are available at probeBase (Greuter et al., 2016). FISH results were positive for EUB (Blue), AMX820 (green) and BFU613 (red), and negative for: Kst157 and Amx1015. The BFU613 was designed for hybridization with "*Ca*. Brocadia fulgida". The TestProbe tool from SILVA (https://www.arb-silva.de/search/testprobe/) shows that this probe can also hybridize with "*Ca*. Brocadia sp.40".

Alignments

hypothetical protein [Candidatus Brocadia sapporoensis]
Sequence ID: **WP_070066018.1** Length: 498 Number of Matches: 2
Range 1: 9 to 497

Score	Expect	Method	Identities	Positives	Gaps	Frame
659 bits(1701)	0.0()	Compositional matrix adjust.	351/489(72%)	389/489(79%)	34/489(6%)	

Features:

```
Query  43   MICAIFFSLIVMISQAGAWDILTLDSNGDPGEFAAITTGTSGASYISHYDDAYGTLKYTT  102
            MICAIFFSL VMISQA AW+ILTLD  GD G+F AITTGTSGASYIS+YD+     LKY T
Sbjct  9    MICAIFFSLAVMISQASAWNILTLDETGDVGQFGAITTGTSGASYISYYDNTNEDLKYAT  68

Query  103  NVSGTWTTITLDSNGNVGRYTSIDVGTSGASYISYYDFTNEDLKYATNVSGTWKYTLVNS  162
            N+SGTWTT TLD+ GNVGR+TSIDVGTSGA YISYYD T+EDLKYATNVSGTW Y +VNS
Sbjct  69   NISGTWTTQTLDNTGNVGRFTSIDVGTSGALYISYYDVTSEDLKYATNVSGTWTYAIVNS  128

Query  163  HNKVGRFSGIKVGTSGAAYISYYYDTAGDLRYATNVSGSWEITTDVDTAGNVGQYASIVV  222
            +N+VGRFSGIKVGTSGAAYIS+YYDTA DLRYATNVSG W +TT+VDTAG+VGQY SIVV
Sbjct  129  NNQVGRFSGIKVGTSGAAYISFYYDTAFDLRYATNVSGLWVVTTNVDTAGSVGQYTSIVV  188

Query  223  GTSGASHISYYDVTNGNLKCATNVSGAWTTETLDSDGDVGQYTSIAMGTSGALYISYYDV  282
            GTSGAS+ISYYD TN +LK AT+ +G WTTETLDS G VGQYTSIAMGTSGALYISYYD
Sbjct  189  GTSGASYISYYDDTNTDLKYATDATGTWTTETLDSTGKVGQYTSIAMGTSGALYISYYDE  248

Query  283  THGDLKYATNVSGTWTTETLDSDGDVGQYTSIAMGTSGASYISYYDVTHGDLKCATNISG  342
            T+G+LK ATN++GTWTT T+DS G+VGQYTSIA+GTSGA YISYYD  + DL+ ATN SG
Sbjct  249  TNGNLKLATNIAGTWTTGTVDSTGNVGQYTSIAVGTSGALYISYYDFDNKDLRYATNASG  308

Query  343  TWTFEAVDSDNKVGINSSISVGLSGDVHICCHDDTDHTLKYASNFE--------------  388
            TWT E VDSD+ VG  SSISVGLSGDV+I +DDT+ LKYA  E
Sbjct  309  TWTTETVDSDDDVGQFSSISVGLSGDVNISYYDDTNKNLKYAFKSETGVPPTPTPTATPT  368

Query  389  --------------------AYEAPPPPPCKVKKLNVKPKSLKLSQGEDAGKTVTLTCKD  428
                                +P P PC  KKLNVKPKSLKLSQGED   KTV LTCKD
Sbjct  369  PGTPTPTPTSTATPTPGTPTPTPSPTPVPCNAKKLNVKPKSLKLSQGEDTEKTVILTCKD  428

Query  429  GFPSVNQPVKAKVMSGKKCVSVSPAEVSTNEDGQATFTITATNKTGTAVVRFKHKNLVGD  488
            GFPSVNQPVKAK++SGKK V+VSPAE  T+EDGQATFTITAT K GTAVVRFKHKNLVGD
Sbjct  429  GFPSVNQPVKAKIVSGKKFVTVSPAEAYTDEDGQATFTITATEKKGTAVVRFKHKNLVGD  488

Query  489  VTVKVKKVS  497
            VTVKVKK +
Sbjct  489  VTVKVKKAT  497
```

Figure S2. Alignment of WP_070066018.1 and WP_070066019.1 using the BLAST tool of NCBI.
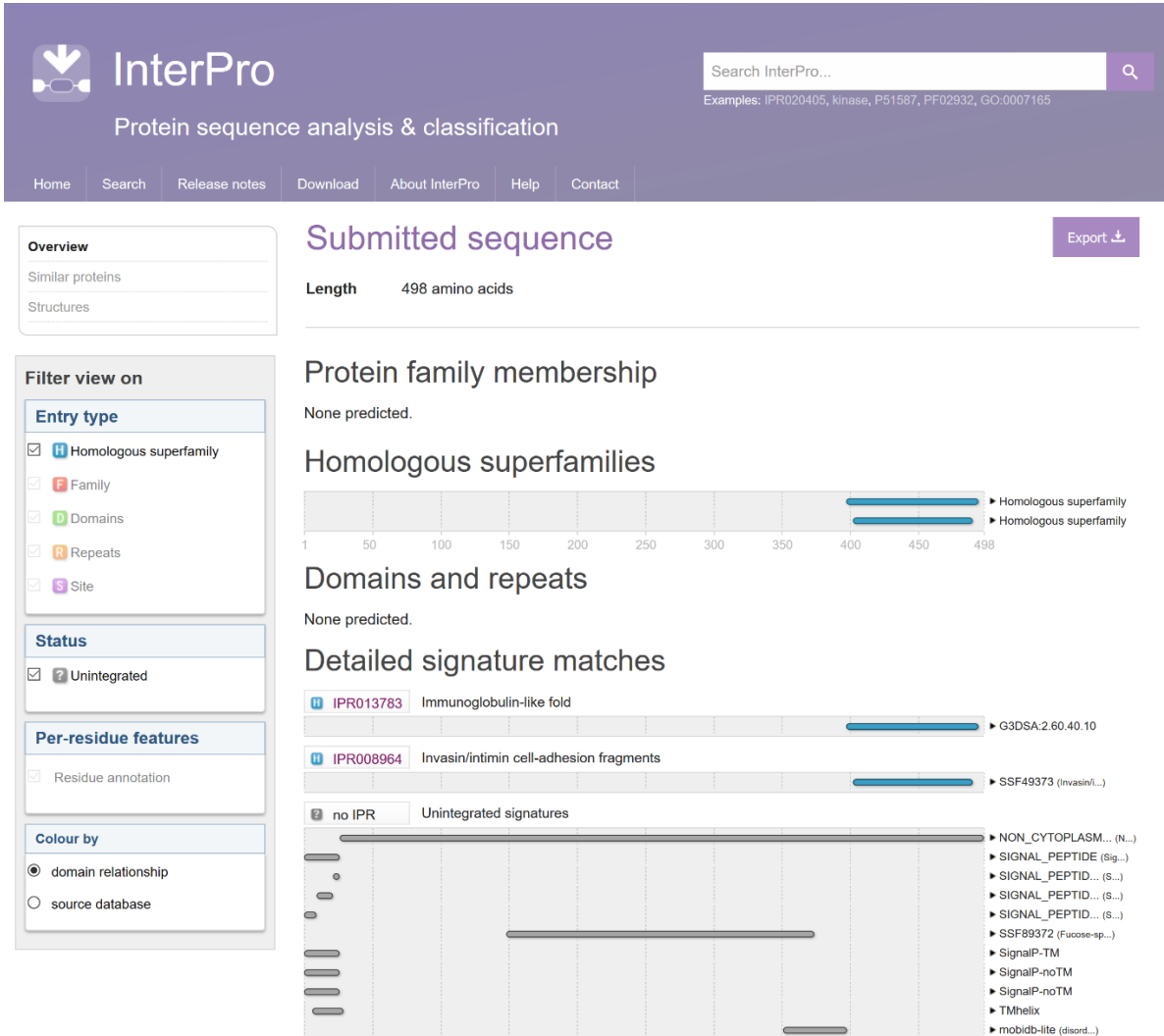
Figure S3. Printscreen of the results of the Interpro protein sequence analysis of WP_07006018.1.
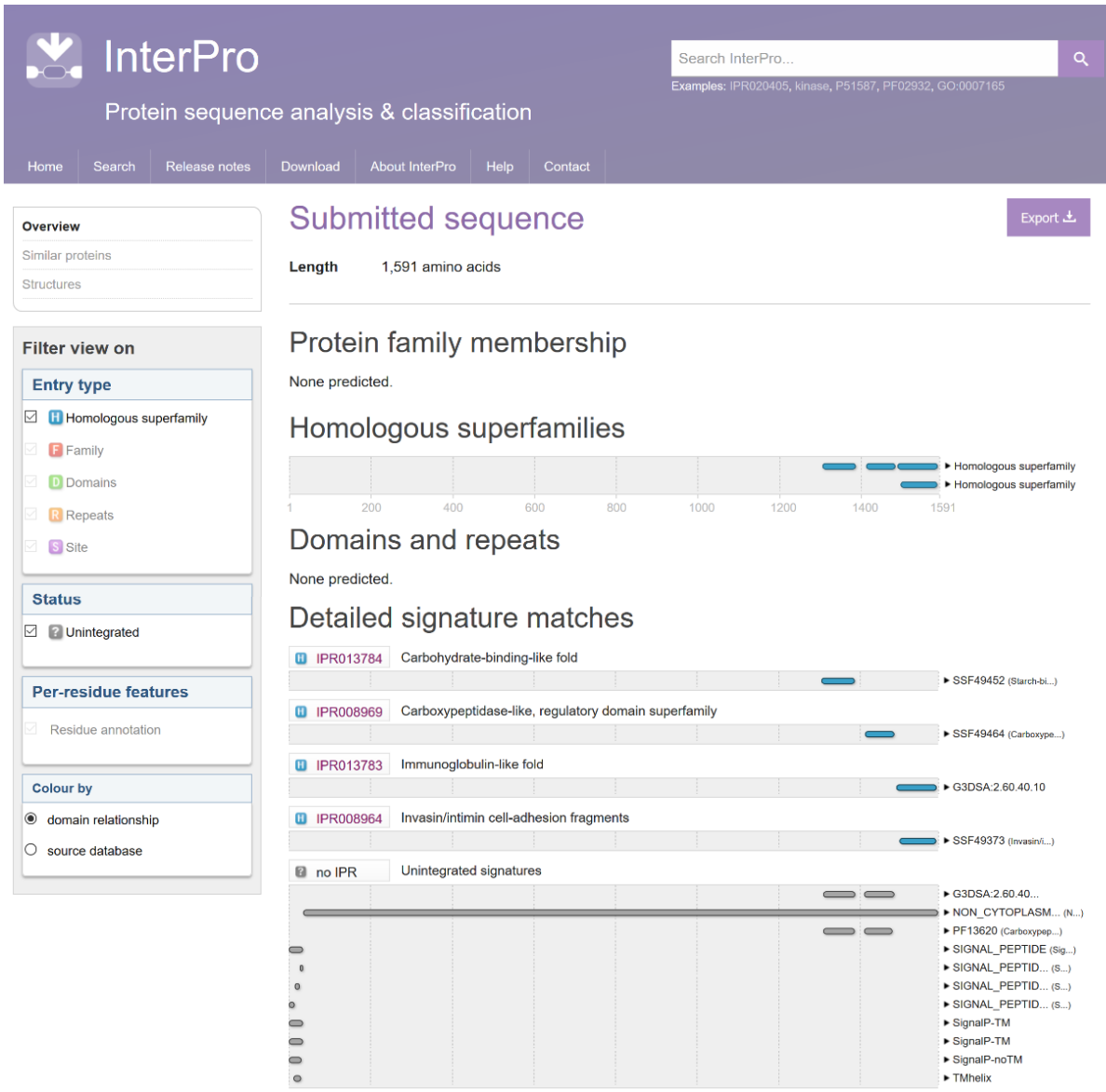
Figure S4. Printscreen of the results of the Interpro protein sequence analysis of S-layer "*Ca*. Kuenenia stuttgartiensis" (SOS03803).

```
>WP_070066018.1 hypothetical protein [Candidatus Brocadia sapporoensis]
MRKKQMGVMICAIFFSLAVMISQASAWNILTLDETGDVGQFGAITTGTSGASYISYYDNTNEDLKYATNI
SGTWTTQTLDNTGNVGRFTSIDVGTSGALYISYYDVTSEDLKYATNVSGTWTYAIVNSNNQVGRFSGIKV
GTSGAAYISFYYDTAFDLRYATNVSGLWVVTTNVDTAGSVGQYTSIVVGTSGASYISYYDDTNTDLKYAT
DATGTWTTETLDSTGKVGQYTSIAMGTSGALYISYYDETNGNLKLATNIAGTWTTGTVDSTGNVGQYTSI
AVGTSGALYISYYDFDNKDLRYATNASGTWTTETVDSDDDVGQFSSISVGLSGDVNISYYDDTNKNLKYA
FKSETGVPPTPTPTATPTPGTPTPTPTSTATPTPGTPTPTPSPTPVPCNAKKLNVKPKSLKLSQGEDTEK
TVILTCKDGFPSVNQPVKAKIVSGKKFVTVSPAEAYTDEDGQATFTITATEKKGTAVVRFKHKNLVGDVT
VKVKKATE


>SOH03803.1 S-layer protein [Candidatus Kuenenia stuttgartiensis]
MKRPGLNLKGNWLSMAGVLFLMLALVMGVASNAKAATGSFDRDRYLPEKAGGNDYDRAWISVTDSSGNTT
SSPDTITVTVRAGVDDATSFILLETGNTSTVFTTTGTVQPNYRALGTTSGYVEDLAGSYWYSVLGNSIVG
LDLRLNAANIGGNAQSATSGDLKVAEGSTLELLYGGSTLDTATVGFNSGSISSSPSSVTSVEGLPSNGNI
ILSISDPDENLNPVVRDVIGFYDGFTTGDTRGTASSMVQISAVDQSSDTGDTLSLGGTSIVARHIMLVET
GNNTGVFTASGYVYGTTTTVAASDLVGNLKVGSSSESYYQGQDITLGAGTVSCTFRIVEANANGGIALIY
NGTSTETYGTVSVVFDATSAPRADVGASLVAFGTSSLSLGDTQTGHKVVCYGTQTSTGYAVSAWGLPSSS
SNYAATGSKLVKLMDGSSYCLIAVSSFSTPNDSAGLGTEGDIIANSGYGSVSVVFDGFRLVGPRDGDTLK
VSYLDLLKSNGQAGTVSANISFGIAGKTGELSVDKTTVDINDFLTVTVIDDNLNTSSSTSQSVGEDSWNG
TTTAERGDRLKVAAYSSGSGEGVGNALSLSFKDGTSVGTSSAIRVSNTDNSLVWIIPKDPSSGYIGTPGT
RSSSAGSATFTLGTQSVDTNALMKGASSAAKSFLSSATTSSFVATLDGLGDTVEISPDGTRWIAIPVTET
GANSATFVGTVGFDYTAARVTTDTTLTRTTAFSDSTDYTTITFANSEFNSTNRLDSFIGTGSVVRVRDDF
SQEFQEVVSVAGSTMVVTKLASSTSFTPYKTWIQVIGNDMTTAREDTLSNGAKVFRIGGFFGATYRIRYN
DAKDADGNYKAGDTLAATASDVGFQTYDGELSTNVSGTIGPNQYVVVTLVDEDLNTSTSSKQTTSEPDRT
ESGTRYVTFYNENGLGNPVGSSTSNAPDGLFKNGFTAKVLYASKLSNVLSSSVDQSSNTIDFKLEETAVN
SGTFKGSFQLSSSSSTLNSSDRLQVSSGDSIYVHYNDSPSATAEDNSTNYRVVGPLSIETSIGSLLLSKA
KAYLSGDTVVVSVVDDDQNTTSGQDTLTDKLKIEGSNYSTGGALTLDLVESGNTSGTFLATFTTGTETVE
TSANLGKVKAQEGGVVTVTYTDSSPSNTTVTETLSFGSYDATIEFNADSYGLGTYAIVTLADAERNTSHT
GTETLLDDVFIQTSSVNSTKMRLVETGNDTGTFVGSILVTSSGATTEFNQIQGAEGETLTVTYSDTINTT
SSTRTVTDTAAITAAVTPSPTPTPGVTPTPGITPTPGLTPTPSLIPGLGSVEGFVTDAATGDGIEGATVR
NQSGIYTDTTDADGFYSIANVEAGTRTFTAVALGYVPSAPTAIVVTAGGTTNLDFALVASVQGTPITTPT
VTPPPTATLVVAVADGDGIPIAGATVTVDGQSGVTDASGGATFTLEAGDYEVSVSATDYLTSVVTVTVTP
PVTILPVTLEPKGPCPEPDEAVASAATVTPDSLSLTKGDSEDVIVLVTDEDGCPAQGVKVKRKLTSANKK
KIKVTPASETTDTSGQATFTVKAKKSKGKANVKFAVKGVKDKPKVNVTLAK
```

Figure S5. The GT motif highlighted in the sequences of the best matched protein and in the S-layer protein of "*Ca*. Kuenenia stuttgartiensis".

Protein

GenPept

This record is a non-redundant protein sequence. Please read more here.

hypothetical protein [Candidatus Brocadia sapporoensis]

NCBI Reference Sequence: WP_070066018.1

Record removed. This protein record was suppressed because it is no longer annotated on any genome

Identical Proteins   FASTA   Graphics

Go to:

```
LOCUS       WP_070066018            498 aa            linear   BCT 11-JAN-2018
DEFINITION  hypothetical protein [Candidatus Brocadia sapporoensis].
ACCESSION   WP_070066018
VERSION     WP_070066018.1
KEYWORDS    RefSeq.
SOURCE      Candidatus Brocadia sapporoensis
  ORGANISM  Candidatus Brocadia sapporoensis
            Bacteria; Planctomycetes; Planctomycetia; Candidatus Brocadiales;
            Candidatus Brocadiaceae; Candidatus Brocadia.
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genomes
            from the same, or different, species.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..498
                     /organism="Candidatus Brocadia sapporoensis"
                     /db_xref="taxon:392547"
     Protein         1..498
                     /product="hypothetical protein"
                     /calculated_mol_wt=52861
ORIGIN
        1 mrkkqmgvmi caiffslavm isqasawnil tldetgdvgq fgaittgtsg asyisyydnt
       61 nedlkyatni sgtwttqtld ntgnvgrfts idvgtsgaly isyydvtsed lkyatnvsgt
      121 wtyaivnsnn qvgrfsgikv gtsgaayisf yydtafdlry atnvsglwvv ttnvdtagsv
      181 gqytsivvgt sgasyisyyd dtntdlkyat datgtwttet ldstgkvgqy tsiamgtsga
      241 lyisyydetn gnlklatnia gtwttgtvds tgnvgqytsi avgtsgalyi syydfdnkdl
      301 ryatnasgtw ttetvdsddd vgqfssisvg lsgdvnisyy ddtnknlkya fksetgvppt
      361 ptptatptpg tptptptsta tptpgtptpt psptpvpcna kklnvkpksl klsqgedtek
      421 tviltckdgf psvnqpvkak ivsgkkfvtv spaeaytded gqatftitat ekkgtavvrf
      481 khknlvgdvt vkvkkate
//
```

Figure S6. Best matched protein (WP_070066018.1). Printscreen of the information of the protein in the NCBI database (https://www.ncbi.nlm.nih.gov/protein/WP_070066018.1?report=fasta).
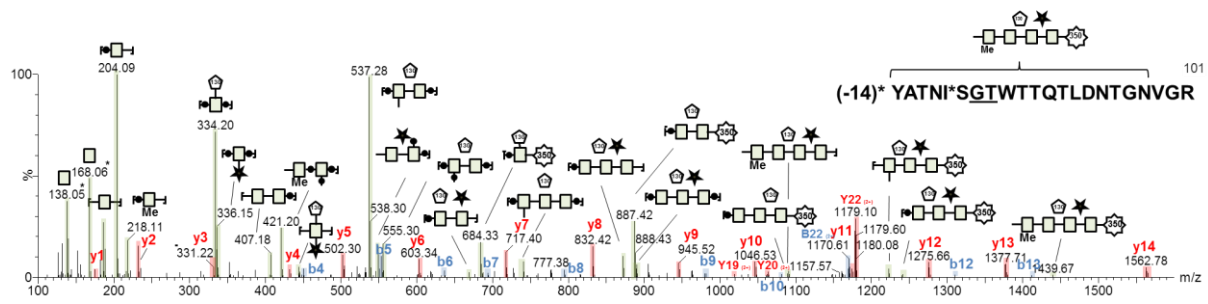
Figure S7. In-gel trypsin digestion and site-specific glycopeptide analysis. Analysis at higher collision energy using a Q-TOF mass spectrometer provides a more complete sequence assignment for both peptide backbone and O-glycan

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	hypothetical protein [Candidatus Brocadia sapporoensis]	660	660	98%	0.0	72%	WP_070066019.1
☐	hypothetical protein B6D34_02820 [Candidatus Brocadia sp. UTAMX1]	650	650	97%	0.0	72%	OQZ04487.1
☐	hypothetical protein A2Y09_06060 [Planctomycetes bacterium GWA2_39_15]	349	1076	96%	5e-108	47%	OHB38428.1
☐	hypothetical protein A3J73_00345 [Planctomycetes bacterium RIFCSPHIGHO2_02_FULL_3	335	577	99%	1e-104	43%	OHB86435.1
☐	hypothetical protein UZ01_01292 [Candidatus Brocadia sinica]	333	1219	76%	2e-103	56%	KXK30623.1
☐	hypothetical protein [Candidatus Brocadia sinica]	332	1220	76%	8e-103	56%	WP_082059331.1
☐	hypothetical protein BROSI_A0470 [Candidatus Brocadia sinica JPN1]	332	1219	76%	9e-103	56%	GAN31966.1
☐	hypothetical protein B6D35_08465 [Candidatus Brocadia sp. UTAMX2]	329	1846	76%	1e-101	56%	OQY99824.1
☐	hypothetical protein B6D34_09185 [Candidatus Brocadia sp. UTAMX1]	328	1384	75%	1e-101	54%	OQZ03019.1
☐	hypothetical protein [Candidatus Brocadia sapporoensis]	320	1966	73%	2e-98	54%	WP_080325039.1
☐	hypothetical protein BROFUL_02012 [Candidatus Brocadia fulgida]	320	1823	76%	2e-98	55%	KKO19281.1
☐	hypothetical protein A2Y09_02495 [Planctomycetes bacterium GWA2_39_15]	314	314	70%	1e-97	49%	OHB35490.1
☐	hypothetical protein AYP45_15535 [Candidatus Brocadia caroliniensis]	318	2802	77%	1e-97	54%	OOP55314.1

Figure S8. Printscreen of the results of the BLAST search with sequence WP_070066018.1 with the BLAST tool from NCBI. The protein was found in different "*Ca*. Brocadia" species with identities between 54% and 72%.