Genomic diversification of giant enteric symbionts reflects host dietary lifestyles

David Kamanda Ngugi[a,1,2], Sou Miyake[a,b,2], Matt Cahill[a], Manikandan Vinu[a], Timothy Hackmann[c], Jochen Blom[d], Matthew D. Tietbohl[a], Michael L. Berumen[a], and Ulrich Stingl[a,e]

[a]Red Sea Research Center, Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia; [b]Temasek Life Sciences Laboratory, 1 Research Link, National University of Singapore, Singapore 117604; [c]Department of Animal Sciences, University of Florida, USA; [d]Bioinformatics and Systems Biology, Justus Liebig University of Giessen, D-35392 Giessen, Germany; and [e]Institute of Food and Agricultural Sciences, Department of Microbiology and Cell Science, University of Florida, USA.

[1] Correspondence to: davkam30@gmail.com

SUPPLEMENTARY INFORMATION

Fish sample collection

This research was undertaken in accordance with the policies and procedures of the King Abdullah University of Science and Technology (KAUST). Permissions relevant for KAUST to undertake the research have been obtained from the applicable governmental agencies in the Kingdom of Saudi Arabia. The Institutional Biosafety and Bioethics Committee at KAUST approved the current research work under permit numbers: 15IBEC31_Stingl and 15IBEC10_Berumen.

Samples for the single-cell genomics and metagenomic datasets of *Acanthurus sohal, Acanthurus nigrofuscus,* and *Naso elegans* were collected as described in Miyake et al.[1] from Abu Shosha reef (22°18'13.74"N, 39° 2'51.79"E), while *Naso unicornis* was collected from Qita Al-Kirsh reef (22°25'50.89"N, 38°59'41.99"E). Single-cell genomics and metagenomics procedures are described separately in the respective subsections below.

Single-cell sorting, genome amplification, and reconstruction

An adapted protocol was used for cell sorting, lysis, and DNA extraction of "*Epulopiscium*"-like cells from the midgut of *Acanthurus nigrofuscus* and *A. sohal*. Briefly, 250 µl of midgut fluid contents were filtered twice through a 40-µm filter to eliminate small-sized cells and algal debris. The retentate was then washed twice with 1 ml of 50% ethanol, after which around 100 large "*Epulopiscium*" cells were picked with a modified glass capillary or pipette under a stereomicroscope, and washed five times; twice in 400 µl of 50% ethanol followed by three washes in 800 µl EB buffer. Approximately 60 cells were then transferred in small volumes and washed again in 400 µl EB. Subsequently, 50 cells were individually picked in 25 ml of EB and placed in a 96-well plate, one cell per well. Several wells were left vacant with 25 ml of EB as control. The 96-well plate was span down to pellet the cells, followed by 5 rounds of freeze (liquid nitrogen) and thaw cycles (65°C) to lyse the cells. The contents of each well were gently mixed by pipetting and 5 µl of each lysate was subsampled for DNA quantification using the QuanTi PicoGreen dsDNA Assay kit (Invitrogen, USA) according to the manufacturer's instructions with standard concentrations ranging

between ~15 pg/µl to 1ng/µl. Wells with high concentrations of genomic DNA were subjected to multiple displacement amplification (MDA) using the Repli-G Midi kit (Qiagen, Netherlands). SYBR Green dsDNA Assay (Invitrogen, USA) was included in the MDA reactions and incubated overnight at 30 ^{o}C in real-time PCR using the CFX Connect detection system (Bio-Rad Laboratories,CA, USA), measuring dsDNA concentrations every 10 minutes for 90 cycles. Multiple controls including positive, negative, multiple cells and wash buffer (from previous isolation of single cells) were included for comparison. The resultant MDA products were then PCR-amplified with universal 16S rRNA gene primers for bacteria (27F/1492R) and also using "*Epulopiscium*"-specific primers (579uF/1232R) following Miyake et al.[2] to further ensure no contaminants were amplified in the wells. Finally, the resultant MDA products—a total of seven independent cells encompassing three different "*Epulopiscium*"-like bacteria—were whole-genome sequenced using the Illumina HiSeq 2000 platform at the KAUST Bioscience Core Laboratory (BCL). Illumina paired-end libraries (2 × 101 bp) were constructed following the standard protocols for genome sequencing and sequenced on a single lane.

The sequenced raw reads were quality-trimmed, while removing adaptor sequences, using Trimmomatic v0.32[3] with the following parameters: ILLUMINACLIP::4:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:60. The "fastx_trimmer" script from the FASTX Toolkit v0.0.13.1[4] was used for an additional quality control to trim either end of the sequences (parameters: -f 20 -m -t 10 -m 60). The phage reads from the internal sequencing standard PhiX 174 were subsequently removed by mapping the quality-trimmed reads against the PhiX 174 genome using Bowtie2 v2.2.4[5] with the following parameters: -q -I 0 --sensitive -t --quiet --qc-filter -X 101. These high-quality sequences were then assembled independently using three *de novo* assembly programs that are suitable for single-cell amplified genome data, namely SPAdes v3.5.0[6], Velvet v0.7.62[7], and IDBA-UD v1.1.1[8]. SPAdes was operated in the single-cell and error-correction modes with the following additional settings: -k 21,33,55,77 --careful --mismatch-correction. IDBA-UD was run with the pre-correction mode; other parameters were --mink 20 --maxk 80 --min_contig 500. Prior to assembly using Velveth-sc, we run the "velvetOptimizer.pl" script (v2.2.5; http://bioinformatics.net.au/software.velvetoptimiser.shtml) to determine the best kmer value with the options –s 33 –e 80 --optFunckmer 'n50'. Subsequently, Velveth-sc was run with the optimised "k" parameter of 77 in the "shortPaired" mode using the error-corrected reads from SPAdes.

The assembled contigs from the above assemblers were subsequently used to generate an integrated set of contigs using CISA v1.3[9], which effectively reduced the number of contigs and increased the N50 size of the assemblies. CISA was run with the following settings: min_length =1000, Gap = 11, R2_Gap=0.99, and genome=3800000, the expected genome size on the basis of data from Mendell et al.[10]. The resulting assemblies were filtered to retain contigs of lengths ≥1000 base pairs, which is within the range of a bacterial gene, followed by fidelity and contamination checks using CheckM v1.0.3[11]. The SAGs were assessed for completeness and contamination based on the presence or absence of 202 *Clostridia*-specific single-copy marker genes using the "taxonomy_wf" command in CheckM. The resultant contigs were scaffolded and gapfilled using the "roundup" workflow of FinishM v0.07 (https://github.com/wwood/finishm) using the

error-corrected paired-end reads from SPAdes. Finally, local misassemblies, small indels and single base differences were identified and polished using Pilon v1.18[12]. Coverage information was obtained by mapping reads to the assembled scaffolds using Bowtie2 v2.2.4[6] with default parameters.

Metagenomic sample preparation, sequencing, and analysis

The extraction of community DNA from the intestine of *Acanthurus sohal, Naso elegans* and *Naso unicornis* was performed using the phenol-chloroform protocol of Ngugi et al.[13] based on ~2 ml of the midgut fluid content. DNA yield ranged from 3–5 µg per sample. The extracted DNA was used to construct whole-genome shotgun (WGS) libraries with the TruSeq library Prep kit (Illumina, San Diego, CA, USA). WGS libraries were sequenced using the Illumina MiSeq (for *A. sohal* and *N. elegans*) or the HiSeq 4000 (for *N. unicornis*) platforms using Illumina paired-end libraries on a 2 × 250 bp or 2 × 150 bp lane, respectively, as part of a larger multiplexed pool at the KAUST BCL. Raw reads were quality trimmed and processed as described above using Trimmomatic[3] and FASTX Toolkit[4], in addition to the removal of internal phage standard reads using Bowtie2[5]. These high-quality sequences were then assembled using SPAdes v3.8.2[6] with the error-correction and metagenomic modes. The WGS contig sequences (filtered to 500 bp) have been deposited in GenBank with the following accession numbers: MDSO00000000 (*A. sohal*), MDSP00000000 (*N. elegans*), and MDSQ00000000 (*N. unicornis*).

To predict protein-coding genes from the size-filtered contigs, we first employed the RNAs prediction workflow of RAMMCAP[14] to mask putative tRNAs and rRNAs in the assembled contigs prior to ORFs prediction using Prodigal v2.6.2[15] in the metagenomic mode. To identify putative carbohydrate-active enzymes (CAZymes) in the intestinal metagenomes of *A. sohal* and the two *Naso* species, the predicted protein-coding genes were searched locally against a CAZyme database that includes glycoside hydrolases (GHs), polysaccharide lyases (PLs), carboxyl esterases (CEs), and glycosyltransferases (GTs) using dbCAN and the accompanying profiled HMMs of the different CAZYmes (http://csbl.bmb.uga.edu/dbCAN/; [16]), and classified into families following the assignment scheme of the CAZYmes database [17]. The proportion of GHs, PLs, CEs and GTs was calculated as the sum of these carbohydrases in each sample.

Recovery and quality assessment of population genomes

Population genomes (PGs) were recovered from the above intestinal metagenomic assemblies of *A. sohal*, *N. elegans*, and *N. unicornis*; these were selected because they harbour different "*Epulopiscium"*-like bacterial clades based on previous 16S rRNA gene diversity studies and FISH-based analyses[2]. The independently assembled contigs (see above) were grouped into genome bins using MetaBAT v0.26.1[18]. MetaBAT integrates empirical probabilistic distances of genome coverage and tetranucleotide signatures to group contigs into putative population genomes. Because our datasets likely includes contig sequences from multiple "*Epulopiscium*"-like clades and probably other diverse prokaryotic communities, we elected to use very conservative binning parameters (p1=97, p2=97, minProb=97, minBinned=30), to minimise contamination and facilitate distinction between closely related "*Epulopiscium*"-like genotypes.

The binned genomes were subsequently improved using two iterative approaches. The first approach employed the workflows implemented in CheckM[11] including: (1)

assessing genome bins for completeness and contamination based on lineage-specific single-copy marker genes (SCMGs) in each bin, and (2) refinement of bins by merging bins with complementary SCMGs and removal of outlying contigs based on the distribution of their GC and tetranucleotide signatures, respectively, using the "merge" and "outliers" commands. Secondly, gaps were filled and scaffolds were formed from the resultant bins using the "roundup" workflow of FinishM, and further polished using Pilon[12] as described above for the SAGs. PGs that were less than 25% complete or with greater than 10% contamination were discarded after assessing *Clostridia*-like bins for completion and contamination based on 202 *Clostridia*-specific SCMGs using the "taxonomy_wf" command in CheckM. Coverage information was obtained by mapping the specific metagenomic reads to the respective PGs using Bowtie2 with default parameters.

Genome annotation, metabolic reconstruction, and comparative genomics

The Rapid Annotation using Subsystem Technology (RAST) pipeline[19] and the NCBI's Prokaryotic Genome Annotation Pipeline[20] were used for all genome annotations. Detailed metabolic reconstruction was done in the KEGG Automatic Annotation Server (KAAS; [21]). Carbohydrate-active enzymes, including glycoside hydrolases, polysaccharide lyases, glycosyltransferases, and carboxyl esterases were identified locally using dbCAN [16], as described in the metagenomic section (see above). SMART was used to predict the modular structures of protein-coding genes (http://smart.embl-heidelberg.de), whereas MAPLE was used to characterise the completeness of KEGG functional modules [21]. Transport proteins were deduced via the web-based transporter (TransAAP) annotation tool (http://www.membranetransport.org/). The theoretical subcellular localisation of the protein-encoding genes was predicted using the PSORTb v3.0 server for determining the subcellular location of proteins[22].

Comparative (pan)-genomics was performed via the EDGAR phylogenomic and comparative analysis web server (https://edgar.computational.bio.uni-giessen.de/; [23]) as described in Ngugi et al.[24,25]. Delineation of closely related pairs of genotypes was done based on the average amino acid identity (AAI) metric[26] and phylogenomic inference using sixteen conserved single-copy genes (**Figure 3**; **Dataset S4**). Briefly, orthologous genes between genome pairs were identified based on a consensus approach employing the OrthoMCL- and the COGtriangles-based clustering methods implemented in the "GET_HOMOLOGUES" software package[27]. The OrthoMCL-based pairwise amino acid identity matrix was then used to plot a heatmap of the AAI (**SI Appendix, Fig. S4**).

For the genome-based phylogenetic inference, we calculated the gene clusters supported by the two clustering methods using the "compare_clusters.pl" script, with the options "–e" and "–t 36" to exclude in-paralogues, and to restrict the output to cover only single-copy genes (SCGs) present in the genomes being compared. A total of sixteen SCGs (**Dataset S4**) were then used for phylogenetic inference. These SCGs were concatenated separately for each genome and aligned using MUSCLE[28] as implemented in Geneious Pro v8.1.4 (http://www.geneious.com) with default settings. The resultant alignment was further manually checked and trimmed using GBLOCKS v0.91b[29] to eliminate poorly aligned positions and highly divergent regions based on the liberal settings that retain the maximum number of characters in the alignment. The final alignment consisted of 2,828 characters with 56% of the position being conserved. A

maximum-likelihood (ML) phylogenetic tree was reconstructed using RAxML v7.2.8[30] as implemented in Geneious Pro under the general time reversible (GTR) + GAMMA model, as determined by ProTest3[31], with 1,000 bootstraps and a maximum-likelihood search of the best tree topology. *Leuconostoc buccalis* was used for rooting the tree.

ML trees for the 16S rRNA gene (**SI Appendix, Fig. S3**) and genes encoding carbohydrases specific for the red-agal sulphated galactans (i.e., β-agarases, β-porphyranases, and carrageenases; **SI Appendix, Fig. S8**) were constructed, respectively, as described in Miyake et al.[2] and Hehemann et al.[32] in Geneious Pro. ML-based phylogenetic inferences for other carbohydrases (alginate lyases and fucoidases; **SI Appendix, Fig. S9**, **S11**, and **12**), targeting polysaccharides of brown algae, were constructed as follows. First, we retrieved protein sequences of experimentally characterized enzymes from Swissprot/Uniprot based on available literature and information from the CAZyme database. Next, putative genes encoding such enzymes in the "*Epulopiscium*" genomes, as deduced using dbCAN, were searched against the non-redundant NCBI database to retrieve closely related homologs. Collectively, these protein sequences were aligned independently for each enzyme using MUSCLE based on default parameters in Geneious Pro. The alignment was then filtered to retain positions that were conserved in 50% of the proteins, followed by phylogenetic inference using RAxML with 1000 bootstraps and the GTR+GAMMA amino acid substitution model. The final filtered alignments used for constructing trees for the respective enzymes consisted of 2,439 (alginate lyases), 1,437 (fucoidases), and 1,898 (endoglucanases) characters.

Metatranscriptomic sample preparation, RNA extraction, sequencing, and analysis

For the metatranscriptome-based temporal studies, we sampled *A. sohal* directly from the Abu Shosha reef (22°18'13.74"N, 39° 2'51.79"E) on 28/04/2013. At least three replicates of *A. sohal* were collected every two hours during the day, and every 4 hours at night starting from 8 am (sampled at 8 am, 10 am, 12 pm, 2 pm, 4 pm, 6 pm, 8 pm, 12 am, and 4 am). Fishes were collected directly by spearfishing during the day (8am – 6pm), but this was not possible at night due to local regulations. Instead, samples sacrificed at night (8 pm – 4 am) were collected well before dusk with a net and were kept in a large enclosure (~3.5 × 2.5 m). Note that the nighttime samples were collected well before sunset to decrease the effect of stress as much as possible. For each individual fish, approximately 500 µl of the midgut content was collected in microcentrifuge tube by making a small incision using a sterile scalpel on the intestinal portion where "*Epulopiscium*"-like bacteria were reported to be highest in abundance[33,34]. Samples were stored at 4 °C overnight in 10× RNAlater (v/v; ThermoFisher Scientific, USA), before the supernatant was removed and the remaining RNA-fixed samples were stored at –80 °C until RNA extraction.

Standard RNA extraction using RNeasy Mini kit (QIAGEN, Netherlands) in itself was ineffective, suggesting that "*Epulopiscium*" cell walls are resistant to enzymatic treatment or that the gut content contained inhibitory compounds. Similarly, the bead lysis method previously employed by Miyake et al. [1] sheared the total RNA, reducing the quality of the RNA yield. Thus, RNA extraction protocol was optimized using freeze-thaw cycles prior to extraction by RNAeasy Mini kit (QIAGEN, Netherlands). The total RNA quality and quantity were measured for once, three, five, and ten times of freeze-thaw cycles with

BioAnalyzer 2100 Total RNA Nano Series II (Agilent Technologies Inc., USA), which revealed three cycles of freeze-thaw to be optimal.

For all samples, the RNAlater-fixed cells were concentrated by centrifugation (8,000 × g for 5 min.), placed in LRT buffer (from RNeasy Mini kit) and subjected to three cycles of freezing (in liquid nitrogen) and thawing. The lysate was then homogenised and total RNA was extracted using the RNeasy Mini kit (QIAGEN, Netherlands) following the standard protocol with DNase treatment (Life Technologies, USA). The total RNA in each sample was purified with the RNeasy MinElute Cleanup kit (QIAGEN, Netherlands). Given our experimental design—in which we intended to use the 16S and 18S rRNA genes for community profiling of the active microbial fraction of the gut—and given that proportions of rRNA relative to total RNA was relatively low (as indicated by BioAnalyzer 2100 Total RNA Nano Series II), we retained all rRNAs present in the total RNA pool of all samples. To compensate for the potential reduction in the mRNA quota for each sample, we increased the overall sequencing depth as summarised in **Dataset S9** (on average 28 ± 14 million reads per sample). Only samples with (i) total RNA concentration of >300 ng μl^{-1}, (ii) 260-to-280 ratio of ~2.1, (iii) 260-to-230 ratio of ~1.4 , and (iv) RNA Integrity Number (RIN) of >8 were RNA-sequenced, an empirical measure for RNA quality[35]. Consequently, all three replicates for the 20:00 time point were discarded, due to low RIN (< 7.5). Library preparation and sequencing was carried out using the Illumina HiSeq 2000 platform at the KAUST BCL. A strand-specific RNA sequencing library was prepared from 2 µg of total RNA per sample using Illumina TruSeq Small RNA Sample Prep Kit. In the end, twenty-seven samples were multiplexed and sequenced using three lanes with 101 cycles.

Raw RNA sequence reads of each sample were quality-trimmed using Trimmomatic v0.3.2[3], using the following parameters: ILLUMINACLIP::4:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:60 and also removing sequencing adaptors. The "fastx_trimmer" script from the FASTX Toolkit v0.0.13.1[4] was used for an additional quality control to trim both sequence ends (parameters: -f 20 -m -t 10 -m 60). Phage PhiX sequences were subsequently removed by mapping the quality-trimmed reads against the PhiX genome using Bowtie2 v2.2.4[5] with the following parameters: -q -I 0 --sensitive -t --quiet --qc-filter -X 101. Potential errors in the reads were then corrected using SPAdes' implementation of the BayesHammer error-correction tool with default parameters[6,36]. The resultant sequences were then partitioned into rRNAs-only reads (including 16S, 23S, 18S, and 5S) and mRNA reads using SortMeRNA v1.99[37]. The raw RNASeq reads have been deposited in the Short Reads Archive under the accession number SRP083815.

Simultaneous transcriptional profiling of host and gut symbionts

To be able to follow the global gene expression of the host and its entire gut microbiota (including *"Epulopiscium"* clade members), we *de novo* assembled each metatranscriptome independently using Trinity v.2.1.1[38]. The assembler was run on the merged, concatenated fastq files of read 1 and 2 of each sample with the following parameters: parameters –seqType fq –single –run_as_paired –min_contig_length 300. **Dataset S10** summarizes the general features and statistics of the assembled transcripts. The resulting transcripts averaged 15,793 ± 9,915 transcripts per sample (n = 27) with mean lengths of 680 ± 101 bp. However, as some assembled transcripts were as long as

32 kbp—that is, longer than the average bacterial coding gene length of ~600 bp—and to confidently annotate the transcripts at the protein level, we next identified candidate-coding regions within each transcript sequence using TransDecoder v2.01 (http://transdecoder.github.io/). TransDecoder identifies coding sequences in the transcript sequences and also gives their matching peptide sequences. The duplicated coding sequences found within larger transcripts were also removed. We maximised the sensitivity of capturing additional open-reading frames (ORFs) with potential functional significance by simultaneously also scanning all ORFs for homology to known proteins in the PFAM database as implemented in TransDecoder.

To evaluate the contribution of gut symbionts (and/or the host) to the digestive processes, we performed a functional analysis of carbohydrate active enzymes (CAZymes) by interrogating the transcripts using dbCAN[16], followed by a taxonomic assignment (described below) of the identified CAZyme families in each metatranscriptome. dbCAN was run locally using the program hmmsearch within HMMER v3.1b2 (hmmer.org) with default parameters based on the provided profile-hidden Markov models (dbCAN HMMs 5.0; http://csbl.bmb.uga.edu/dbCAN/download.php) as described above.

Taxonomic assignment of coding genes and transcripts

The taxonomic assignment of all metagenomic protein-coding genes and the nearly full-length transcripts, including the predicted carbohydrases, was performed to evaluate the relative contributions of the host and the different symbionts to metabolic functions in the gut. For taxonomic profiling, protein-coding genes and transcripts were queried using DIAMOND v0.8.8 [39] with default settings and an e-value of 10^{-5} against the NCBI's non-redundant protein database (downloaded on 01.06.2015). The resulting blast output was filtered to retain matches to prokaryotic and eukaryotic genes above a bit score of 50 and with alignment coverage of 80%. The gi number to taxonomy mapping file provided by NCBI (gi_taxid_prot.dmp), in conjunction with the NCBI taxonomy node hierarchy file (nodes.dmp), was used to gather full taxonomic information of NCBI's classification, and consolidated into different taxonomic rankings using the scripts provided in the GenBank taxonomy processing tools repository (https://github.com/spond/gb_taxonomy_tools). Finally, the relative abundance of each phyla or genera in a given sample was calculated relative to the total number of transcripts in each sample.

At the time of our analysis, the NCBI database contained only the genomic data from the draft genome of "*Epulopiscium*" species type B from *Naso tonganus*[40], whose metabolism has not been formerly described. Alongside the phylogenomic divergence of genotype A1 and J/C relative to the A2/B genotypes (**SI Appendix, Fig. S3 and S4**), we were prompted to examine whether our ability to classify is hampered by the paucity of distantly related reference genomes in the NCBI database. To this end, we performed classification based on a customized database comprising the non-redundant database and the genomic sequences of all SAGs and PGs generated in this study, and compared these results to those based on the original database (**SI Appendix, Fig. S2**). The taxonomic assignment of all coding genes (**SI Appendix, Fig. S2a**) and glycoside hydrolases (GHs; **SI Appendix, Fig. S2b**) in each metagenome against the non-redundant database with and without the fourteen genomes revealed that there was a significant improvement in phylogenetic anchoring of "*Epulopiscium*"-like coding genes and GHs by approximately

21–60%. The greatest increase is observed for the *Naso* species presumably due to the fact that they are predominantly colonized by "*Ca.* Parepulopiscium" species types J and C[2], which are divergent to both "*Ca.* Epulopisciides" species types A2/B and "*Ca.* Epulopiscium fishelsoni" type A1 (**SI Appendix, Fig. S4**). This increase is also reflected in the overall alignment coverage of 88–95% (**SI Appendix, Fig. S2c**) and their corresponding high average amino acid identities (90–91%) of the proteins (**SI Appendix, Fig. S2d**). Accordingly, subsequent taxonomic assignments of the coding genes (in the metagenomes) and metatranscripts were conducted based on the customized database.

Taxonomic assignment of 16S rRNA transcripts

Prior to taxonomic assignment of rRNA transcripts, sequences that had the following properties were discarded: less than 100 bp in length, with more than six polymers, and with ambiguous bases. The paired-end 16S (both Bacteria and Archaea) and 18S rRNA reads of each dataset—after the SortMeRNA step (see above), were first merged using Pandaseq v2.8[41] (with the parameters: -N –t 0.32), and subsequently classified using MOTHUR v1.37[42]. The resulting non-redundant sequences were taxonomically assigned based on the SILVA v123 taxonomy[43] and a confidence threshold of 80% using the latest 16S and 18S databases provided in the MOTHUR website (http://www.mothur.org/wiki/Silva_reference_files).

To robustly capture the potentially active "*Epulopiscium*" clade members in the temporal metranscriptome, we also selected all MOTHUR-classified "*Epulopiscium*"-like 16S rRNA transcripts and assigned them into specific subclades (or types) using Pplacer v1.1alpha18[44] (with default settings except option "--group 5" as suggested in the accompanying manual). Pplacer places reads onto a reference phylogeny by maximising the phylogenetic likelihood or posterior probability. For our work, we used the maximum-likelihood tree of full-length 16S rRNA genes from Miyake et al.[2]—created using GTR + I + gamma substitution model with 1000 bootstraps—as the scaffold tree upon which the 16S transcripts were phylogenetically assigned to specific clades. Taxonomically assigned rRNAs were then averaged per time point for comparison.

The subclade-specific taxonomic assignment of "*Epulopiscium*"-only 16S transcripts in gut metatranscriptomes from *A. sohal* (**SI Appendix, Fig. S14a**) revealed "*Ca.* Epulopisciides saccharus" type B as the most abundant genotype (44 ± 18%), followed by "*Ca.* Epulopisciides gigas" type A2 (15 ± 15%), "*Ca.* Epulopiscium fishelsoni" type A1 (10 ± 6%), and genotypes RS01 (4 ± 3%), "*Ca.* Parepulopiscium" type C (4 ± 1%), RS03 (3 ± 2%), and "*Ca.* Parepulopiscium" type J (~1%). To exclude the potential effect of multiple rRNA operons in "*Ca.* Epulopiscium" and related giant bacteria, we were prompted to examine in detail the temporal gene expression of the predominant genotypes in *A. sohal* (types A1, A2, and B). To this end, we compared the temporal expression of 16S rRNAs and two housekeeping genes (*oriC* and *ftsZ*), by mapping them against RNASeq reads of 16S rRNAs- and mRNAs only, respectively. Here also, the expression of these genes was consistently higher for the genotype B (and A2; **SI Appendix, Fig. S14b**); however, the expression of the 16S genes was remarkably higher than these single-copy genes. If rRNA operon copy number reflects the ecological strategies of bacteria[45], then it appears that the B and A2 cell types have the ability to

respond rapidly to nutrient accretion based on their rRNAs expression pattern and the peak expression time (early morning).

Genotype-specific gene expression analyses

The paired-end mRNAs reads were mapped to three genomes representing different genotypes of "*Epulopiscium*" clade members, namely "*Ca.* Epulopiscium fishelsoni" type A1 (SCG-B11WGA), "*Ca.* Epulopisciides gigas" type A2 (SCG-C07WGA), and "*Ca.* Epulopisciides saccharus" type B (PG-AS2M_MBin01) using Bowtie2[5] with zero mismatches allowed per alignment. The resulting mapped read counts for each genome were used as input to the R package DESeq2[46], where normalization across libraries was done using the regularized logarithm transformation (rlog) procedure prior to visualization of clustering. Irrespective of the genotype, the samples grouped into clusters dependent on the time of sampling—that is, morning (8 and 10 am), afternoon (12, 2, and 6 pm), and nighttime (1 and 4 am) intervals (**SI Appendix, Fig. S14c** and **S14b**). Raw counts and discrete distributions were used for differential gene expression (DGE) analysis. Statistically significant differentially expressed genes were defined as those with a p-value ≤ 0.05 and a fold-change of ≥ 2 for top hits.

The differential expression of all coding genes in each genotype showed that they not only clustered by time of day (**SI Appendix, Fig. S15**), but also revealed that many were significantly upregulated (a fold-change of ≥ 2 and a p-value ≤ 0.05) in the case of genotype B (34 to 54% of 2,165 genes) and less for genotypes A2 (12–19% of 1,973 genes) and A1 (17–36% of 2,718 genes; **SI Appendix, Fig. S14c**). These results demonstrate that cells related with *Ca.* Epulopisciides saccharus type B are transcriptionally the most active microorganisms in the midgut of *A. sohal*.

Statistical analysis

Statistical analyses were performed using GraphPad Prism v7.0a (GraphPad Software, In.). Significant differences between sample means were tested by conducting an analysis of variance (one-way ANOVA) and correcting for multiple comparisons by controlling the false discovery rate ($\alpha = 0.05$) using the recommended two-stage step up method of Benjamini et al.[47]. These tests were done for the genome-size scaled counts of CAZymes families and COG counts of "*Epulopiscium*" genomes against those of *Clostridium lentocellum*. Polarhistograms and heatmaps were generated using the "phenotypicForest" (http://chrisladroue.com/phorest/) and the heatmap.2 in the "ggplot2" R packages respectively ([48]; https://www.r-project.org/).

DAPI staining for cell counts

4',6-diamidino-2-phenylindole (DAPI) staining for cell counts on different host fish was prepared as described in Miyake et al. [2], which is a modified protocol from Daims et al. [49]. Briefly, gut fluid from the midgut section where high "*Epulopiscium"* abundance has been reported [34] was collected from *A. nigrofuscus*, *A. sohal*, *N. elegans* and *N. unicornis* —all collected in the late morning—and fixed in 4% formaldehyde for a few hours at 4°C. For each fixed sample, a few drops of the fixed sample was placed on microscope slides and mounted in VECTASHIELD Antifade Mounting Medium with DAPI (H-1200, Vector Laboratories, USA). The slides were incubated in the dark for 10 minutes before analysis with fluorescence microscopy using a Zeiss LSM 710 upright confocal microscope at a magnification of 40×. Counts of DAPI-stained "*Epulopiscium"*-like

giant cells (>100μm), "*Epulopiscium"*-like small cells and other Prokaryotic cells were recorded at 10 randomly selected spots on the slide (**Dataset S2**).

Supplementary results/discussion

Proposed classification of novel members of the "Epulopiscium" clade

"Candidatus Epulopiscium fishelsoni" (gen. nov. sp. nov.)

The nomenclature follows the original proposed description by Montgomery and Pollack [50]. "*Epulopiscium fishelsoni"* (L. n. *epulum*, "guest at a banquet"; L. adj. *piscium*, "of a fish", referring to the fact that the organisms were found inside the intestine of a fish; *fishelsoni*, named after Prof. Lev Fishelson, one of the discoverers. This organism encompasses the type A1 subclade represented by four single-cell genomes (SCG-B05WGA; SCG-B11WGA; SCG-C06WGA; SCG-D08WGA) and one population genome (PG-AS2M_MBin02) from *Acanthurus sohal* and *Acanthurus nigrofuscus*.

"Candidatus Epulopisciides gigas" (gen. nov. sp. nov.)

"Epulopisciides gigas" (L. n. *epulum*, "guest at a banquet"; L. gen. *pisci*, "a fish", referring to the fact that the organisms were found inside the intestine of a fish; Suff. *ides*, ending means literally "son of" or "descendant of", used here to describe resemblance to "*Epulopiscium fishelsoni"*; L. gen. *gigas*, "giant", in reference to their enormous cell sizes (~600 μm), which is larger than any of the enteric prokaryotic symbionts. This organism encompasses the type A2 subclade represented by two single-cell genomes (SCG-B10WGA; SCG-C06WGA) from *Acanthurus nigrofuscus.*

"Candidatus Epulopisciides saccharus" (gen. nov. sp. nov.)

"Epulopisciides gigas" (L. n. *epulum*, "guest at a banquet"; L. gen. *pisci*, "a fish", referring to the fact that the organisms were found inside the intestine of a fish; Suff. *ides*, ending means literally "son of" or "descendant of", used here to describe resemblance to "*Epulopiscium fishelsoni"*; L. gen. *saccharus*, "sugars", describes their preference for sugars as inferred from the high density of glycoside hydrolases in their genomes. This organism encompasses the type B subclade and is represented by one single-cell genome (SCG-B10WGA; SCG-C06WGA) and one population genome (PG-AS2M_MBin01) from *Acanthurus nigrofuscus* and *Acanthurus sohal,* respectively.

"Candidatus Parepulopiscium" (gen. nov.)

The description is the same as for the genus *Epulopiscium*. Prefix *par*, "outside of" or "abnormal", referring to their inferred distinctive physiological traits relative to the genus "*Epulopiscium"* and varied host preference. This organism encompasses the types C and J subclades, represented by four near-complete population genomes from *Naso elegans* (PG-Nele67M_MBin01; PG-Nele67M_MBin02; PG-Nele67M_MBin03) and *Naso unicornis* (PG-Nuni2H_MBin01; PG-Nun2H_MBin03).

Supplementary figures

Fig. S1. Schematic representation of the gut of *Acanthurus sohal* in its unraveled state, showing the "*Epulopiscium*"-rich gut section that we sampled; also representative of the other fishes used in this study. Image is not drawn to scale.

Fig. S2. Phylogenetic anchoring of metagenomic data using a customized non-redundant (NCBI) protein database comprising the fourteen genomes generated in this study. (**a**) The relative abundance of "Epulopiscium"-like genes in the metagenomes. Taxonomic assignment was done independently following the NCBI taxonomy with the non-redundant protein database with (cDB) and without (nDB) the fourteen "Epulopiscium" genome sequences. The total number of predicted proteins in each metagenome was 12,467 (*Acanthurus sohal*), 28,187 (*Naso elegans*), and 95,046 (*Naso unicornis*). (**b**) The relative abundance of "Epulopiscium"-like glycoside hydrolases in the metagenomes deduced using the same method as the total predicted proteins. The total counts of predicted GHs are 626 (*A. sohal*), 105 (*N. elegans*), and 106 (*N. unicornis*). (**c**) The overall protein alignment coverage of "*Epulopiscium*"-like GHs and (**d**) the average amino identity over the aligned regions based on the customized database. The bottom and top of the box (in **c** and **d**) indicate, respectively, the first and third interquartiles, while the inside line and plus sign denote the median and mean counts, respectively. The whiskers are located at 1.5 the interquartile range above and below the box.

Fig. S3. Phylogenetic placement of "*Ca.* Epulopiscium fishelsoni" and related giant bacteria (in red) based on 16S rRNA genes. The maximum-likelihood tree was constructed as described in Miyake et al.[2]. Only full-length sequences were used in addition to following the published clade affiliation nomenclature.

Fig. S4. Pairwise average amino acid identity (AAI) of "*Ca.* Epulopiscium" and related giant bacteria based on their orthologous genes. The dendogram on the left shows hierarchical clustering of genomes based on the Bray-Curtis dissimilarity matrix using the average linkage method. Parentheses show clade affiliation of each genotype (see **Figure 3** and **SI Appendix, Fig. S3**). The publicly available draft genome of "*Epulopiscium*" sp. type B (from *Naso tonganus*) is indicated with an asterisk. SCG, single-cell genome; PG, population genomes.

Fig. S5. Distribution of genes encoded in *"Ca.* Epulopiscium" and related giant bacteria based on COG functional categories. Significantly enriched COGs categories are depicted with letters on top of the barcharts; those followed by the same letters are significantly different based on one-way ANOVA of their means ($P < 0.05$). Barcharts indicate the mean (± SD) for types A1 ($n = 5$), A2 ($n = 2$), B ($n = 2$), C ($n = 2$), and J ($n = 3$).

Fig. S6. Genome-size scaled counts of carbohydrases in "*Ca.* Epulopiscium" and related giant bacteria relative to reference biopolymer-degrading bacteria. Boxplots represent genome-size scaled counts of carbohydrate esterases (upper panel), glycosyl transferases (middle panel), and polysaccharide lyases (lower panel) in the major "*Epulopiscium*" genotypes (A1, A2/B, and J/C) and in relation to other *Clostridia* or free-living agarolytic bacteria. Significant differences were measured using one-way ANOVA (** $P < 0.01$; *** $P < 0.001$).

Fig. S7. Polar histogram showing the relative abundance of CAZyme families in the reconstructed "*Ca.* Epulopiscium", related giant bacteria, and reference genomes belonging to agarolytic bacteria and plant biomass-degrading *Clostridia*.

Fig. S8. Maximum-likelihood tree of carbohydrases that target sulfated galactans of red algae in the reconstructed *"Epulopiscium*" genomes. The phylogenetic tree was constructed using following the approach and the reference sequences from Hehemann et al.[32]. Note that majority of the enzymes are encoded in the genomes of "*Ca*. Epulopisciides" species types A2 and B. Proteins of characterized reference organisms are shown with a black circle.

Fig. S9. Unrooted phylogenetic tree of characterized GH5 and GH74 endoglucanases and endoglucanase-like proteins from the *"Epulopiscium*" genomes using maximum-likelihood analysis. Endoglucanases lyases target the highly branched xyloglucan of cholorophycean green algae consisting of β-1,4-glucosyl

residues carrying side chains decorated with α-1,6-xylosyl residues, which are additionally substituted with a β-1,2-galactosyl residue (details in **Figure 1**). Nearly of all endoglucanase-like proteins were found in the genomes of "*Ca*. Epulopicisciides gigas" type A2 and "*Ca*. Epulopicisciides saccharus" types B. Experimentally characterized enzymes are shown with a black circle. The predicted subcellular localization of the all proteins is indicated as well.

Fig. S10. Sequence alignment of characterized GH74 endoglucanases and endoglucanase-like proteins from the "*Ca*. Epulopicisciides" genomes using MUSCLE. The amino acid sequences of catalytic domains of experimentally characterized endo-xyloglucanases are indicated at positions W93, W96, W350, and W351 of *Paenibacillus* sp. KM21 (based on [51]), with conserved tyrosine residues highlighted in purple. Conserved amino acid sequences found in validated enzymes and those occurring in "*Ca*. Epulopicisciides" genomes are indicated with asterisks, while the conserved inserted sequence only found in GH74 endoglucanase-like proteins from "*Ca*. Epulopicisciides" genomes is shown with a grey bar. The active-site blocking extra loop (G397–H400 based on *Paenibacillus* sp. KM21 positions) responsible for the exo-acitivity in *Geotrichum* sp. M128 [52]—the difference between endo-processive and exo-active xyloglucanases—is depicted with a white box. Other details are as indicated in **SI Appendix, Fig. S9**.

Fig. S11. Unrooted phylogenetic tree of characterized GH29 and GH95 α-1,2-fucoidases and fucoidase-like proteins from *"Epulopiscium*" genomes using maximum-likelihood analysis. Proteins of characterized reference organisms are shown with a black circle; the predicted subcellular localization of the all proteins is indicated as well. Note that none of "*Epulopiscium*" genomes encodes the family 95 fucoidases, while all the putative GH29 fucoidases in their genomes lack signal peptides and carbohydrate-binding modules.

Fig. S12. Maximum-likelihood tree of characterized alginate lyases and putative alginate lyase-like proteins from the *"Epulopiscium*" genomes. Alginate lyases target alginate, a unique structural (linear) polysaccharide of brown algae, consisting of two uronic acids—α-L-guluronate and β-D-mannuronate. Note that all the enzymes were only found in "*Ca*. Parepulopicium" species types C and J. Experimentally characterized enzymes are shown with a black circle. Genes encoding characterized heparinases were used for outgroup.

Fig. S13. Genome-size scaled counts of peptidases in "*Ca.* Epulopiscium" and related giant bacteria relative to reference biopolymer-degrading bacteria. Boxplots represent genome-size scaled counts all peptidases (upper panel), serine peptidases (middle panel), and metalloproteases (lower panel) between the major "*Epulopiscium*" clades (A1, A2/B, and J/C) and in relation to other *Clostridia* or free-living agarolytic bacteria. Note that the A1 and J/C genotypes (and other *Clostridia*) are enriched relative to the A2/B genotypes. Additional details are provided in **Datasets S7** and **S8**. Significant differences were measured using one-way ANOVA (*$P < 0.05$; ** $P < 0.01$; ***$P < 0.001$).

Fig. S14. Genotype-specific gene expression patterns within major "*Epulopiscium*" clade members. (**a**) Relative abundance of *"Epulopiscium*" subclades based on 16S transcripts in the gut of *A. sohal*. (**b**) Temporal expression of 16S, OriC, and FtsZ genes in three of the most abundant genotypes. Bars depict the min and max gene expression levels, with means indicated as a solid horizontal line. The grey lines demarcate the average expression of each gene (15.8, 3.2, and 4.2, respectively) for all sampling points in genotype B. (**c**) The proportion of genes in the three genotypes that were significantly up- or down-regulated at three time intervals: morning and afternoon (MvsA), morning and night (MvsN), and afternoon and night (AvsN). Significantly expressed genes are considered as those with an adjusted *p*-value of ≤ 0.05 and a fold change of ≥ 2.

Fig. S15. Clustering of metatranscriptomes by the time of sampling independent of the genotype. Principle component analysis (PCA) was done using regularized logarithm transformed counts of mapped reads as determined using DESeq2.

Fig. S16. Predicted energetics of glucose and citrate fermentation in *"Epulopiscium*" clade members. (**a**) Fermentation pathways for glucose. Lactate can be formed also through an incomplete methylglyoxal shunt, yielding –2 ATP/glucose (not shown). (**b**) F-type ATP synthase and RNF complex driven by redox cofactors generated during glucose fermentation. *n* denotes number of ATP formed by electron transport phosphorylation (ETP; see next panel). If *n* is negative, the ATPase and RNF complex operate in the direction reverse to that illustrated. (**c**) ATP yield from glucose fermentation. (**d**) Citrate transport and fermentation. The translocation of electrons (e^-) out of the cell by CitS is a formalism to make net transport

of citrate by CitS electroneutral; see Dimroth et al.[53]. Other abbreviations: SLP, substrate level phosphorylation; Fd_{red}, reduced ferredoxin; Fd_{ox}, oxidized ferredoxin.

Fig. S17. Multiple sequence alignment of subunit c of the Na^+-specific F-ATP synthase of three "*Ca.* Epulopiscium" and one "*Ca*. Parepulopiscium" species (in bold) in comparison to *Acetobacterium woodii* and *Propionigenium modestum*. The sodium-binding motif is in red, underlined text. Subunit c of the H^+-specific F-ATPase of *Escherichia coli* is included for comparison and does not possess this motif. Residues in binding motif are those indicated in Rahlfs et al.[54], excluding P25. Ten of the fourteen "*Epulopiscium*" genomes do not encode a subunit c gene and are not included in the figure. Locus tags (for "*Epulopiscium*") and GenBank accession numbers (for reference genomes) are shown in brackets.

Fig. S18. Inventory of genes encoding enzymes catalysing membrane energetics (squared symbols) and citrate fermentation (starred symbols) in the "*Epulopiscium*" clade and reference *Clostridia*. Square Symbols indicate genes that are present, either orthologous (solid symbols) or non-orthologous (open symbols) relative to the reference *Clostridia* genomes. The phylogenetic affiliation of all genomes is highlighted on the right side (as in **Figure 1**). For the V-type ATPase, only the subunits C to G were considered as diagnostic of this ATPase type as suggested by Lolkema et al.[55]. If more than two of the six subunits of the RNF complex were present in "*Ca.* Epulopiscium" and related giant bacteria, then they were scored as potentially having the complex. Note that the V-type ATPAse and the soluble [FeFe]-hydrogenase (only one copy) of "*Ca*. Epulopiscium" are non-orthologous to most *Clostridia*.

Dataset S1. A compilation of algae types/species previously found in the stomach contents of herbivorous surgeonfishes.

Dataset S2. A compilation of algae types/species previously found in the stomach contents of herbivorous surgeonfishes.

Dataset S3. Extended stats of the assembled genomes.

Dataset S4. List of single-copy genes (SCGs) used for genome-based phylogenetic inference.

Dataset S5. Distribution of glycoside hydrolases (GHs) putatively targeting plant and algal polysaccharides encoded in "*Ca.* Epulopiscium" genomes, related giant bacteria and *Clostridia* species.

Dataset S6. Distribution of carbohydrate-active enzymes (CAZymes) in "*Ca.* Epulopiscium", related giant bacteria, and reference *Clostridia* genomes, including their counts scaled to the genome sizes (at the bottom of the table).

Dataset S7. Absolute counts and genome-size scaled abundances of predicted peptidases in "*Ca.* Epulopiscium" and related giant bacteria relative to other *Clostridia*.

Dataset S8. Overall counts of predicted serine peptidases and their abundances scaled to the genome sizes in in "*Ca.* Epulopiscium" and related giant bacteria.

Dataset S9. General features of metatranscriptomic datasets generated in this study and the taxonomic assignment of the 16S and 18S transcripts in each sample.

Dataset S10. General features of assembled metatranscriptomic contigs and the putative taxonomic origin of mRNA and glycoside hydrolase transcripts.

Dataset S11. Relative abundance of putative "host-associated" mRNA transcripts in the metatranscriptomes.

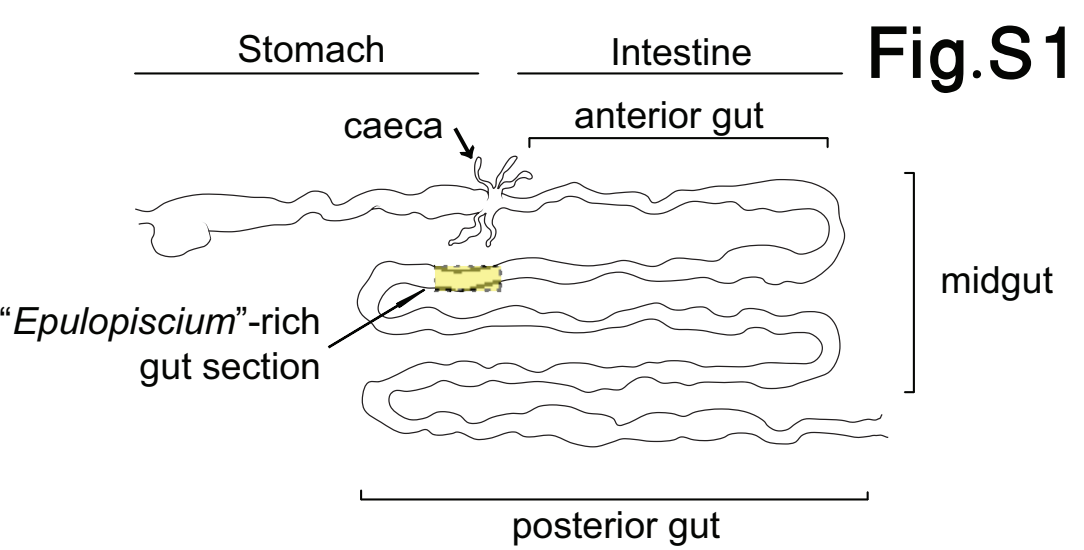
Reference:

1. Miyake, S., Ngugi, D. K. & Stingl, U. Diet strongly influences the gut microbiota of surgeonfishes. *Mol. Ecol.* **24,** 656–672 (2015).
2. Miyake, S., Ngugi, D. K. & Stingl, U. Phylogenetic diversity, distribution, and cophylogeny of giant Bacteria (*Epulopiscium*) with their surgeonfish hosts in the Red Sea. *Front Microbiol* **7,** 285 (2016).
3. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
4. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing

data. *PLoS ONE* **7,** e30619 (2012).
5. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
6. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
7. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18,** 821–829 (2008).
8. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).
9. Lin, S.-H. & Liao, Y.-C. CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes. *PLoS ONE* **8,** e60843 (2013).
10. Mendell, J. E., Clements, K. D., Choat, J. H. & Angert, E. R. Extreme polyploidy in a large bacterium. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 6730–6734 (2008).
11. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25,** 1043–1055 (2015).
12. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* **9,** (2014).
13. Ngugi, D. K., Antunes, A., Brune, A. & Stingl, U. Biogeography of pelagic bacterioplankton across an antagonistic temperature-salinity gradient in the Red Sea. *Mol. Ecol.* **21,** 388–405 (2012).
14. Li, W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* **10,** 359 (2009).
15. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
16. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **40,** W445–51 (2012).
17. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* **42,** D490–5 (2014).
18. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3,** e1165–e1165 (2014).
19. Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9,** 75 (2008).
20. Angiuoli, S. V. *et al.* Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. *OMICS* **12,** 137–141 (2008).
21. Takami, H. *et al.* Evaluation method for the potential functionome harbored in the genome and metagenome. *BMC Genomics* **13,** 699 (2012).
22. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *J Gerontol* **26,** 1608–1615 (2010).
23. Blom, J. *et al.* EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* **10,** 154 (2009).
24. Ngugi, D. K. *et al.* Comparative genomics reveals adaptations of a halotolerant thaumarchaeon in the interfaces of brine pools in the Red Sea. *The ISME Journal* **9,** 396–411 (2015).
25. Ngugi, D. K., Blom, J., Stepanauskas, R. & Stingl, U. Diversification and niche adaptations of Nitrospina-like bacteria in the polyextreme interfaces of Red Sea brines. *The ISME Journal* (2015). doi:10.1038/ismej.2015.214
26. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology* **10,** 504–509 (2007).
27. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* **79,** 7696–7701 (2013).
28. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797 (2004).
29. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and

ambiguously aligned blocks from protein sequence alignments. *Systematic Biol.* **56,** 564–577 (2007).
30. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biol.* **57,** 758–771 (2008).
31. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164–1165 (2011).
32. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464,** 908–912 (2010).
33. Fishelson, L., Montgomery, W. L. & Myrberg, A. A. A unique symbiosis in the gut of tropical herbivorous surgeonfish (Acanthuridae, Teleostei) from the Red Sea. *Science* **229,** 49–51 (1985).
34. Clements, K. D., Sutton, D. C. & Choat, J. H. Occurrence and characteristics of unusual protistan symbionts from surgeonfishes (*Acanthuridae*) of the Great Barrier Reef, Australia. *Marine Biology* **102,** 403–412 (1989).
35. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7,** 3 (2006).
36. Nikolenko, S. I., Korobeynikov, A. I. & Alekseyev, M. A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14 Suppl 1,** S7 (2013).
37. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts611
38. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8,** 1494–1512 (2013).
39. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12,** 59–60 (2014).
40. Miller, D. A., Suen, G., Clements, K. D. & Angert, E. R. The genomic basis for the evolution of a novel form of cellular reproduction in the bacterium *Epulopiscium*. *BMC Genomics* **13,** 265 (2012).
41. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13,** 31 (2012).
42. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75,** 7537–7541 (2009).
43. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41,** D590–D596 (2012).
44. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11,** 538 (2010).
45. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology* **66,** 1328–1333 (2000).
46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).
47. Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93,** 491–507 (2006).
48. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis - Hadley Wickham - Google Books*. (Springer-Verlag, 2009). doi:10.1007/978-0-387-98141-3
49. Daims, H., Stoecker, K. & Wagner, M. in *Molecular Microbial Ecology* (eds. Osborn, A. M. & Smith, J. C.) 192–211 (Molecular microbial ecology, 2005).
50. Montgomery, W. L. & Pollak, P. E. *Epulopiscium fishelsoni* N. G., N. Sp., a protist of uncertain taxonomic affinities from the gut of an herbivorous reef fish. *The Journal of Protozoology* **35,** 565–569 (1988).
51. Matsuzawa, T., Saito, Y. & Yaoi, K. Key amino acid residues for the endo-processive activity of GH74 xyloglucanase. *FEBS Lett* **588,** 1731–1738 (2014).
52. Yaoi, K. *et al.* The structural basis for the exo-mode of action in GH74 oligoxyloglucan reducing end-specific cellobiohydrolase. *J Mol Biol* **370,** 53–62 (2007).
53. Dimroth, P. Primary sodium ion translocating enzymes. *Biochim. Biophys. Acta* **1318,** 11–51 (1997).
54. Rahlfs, S. & Müller, V. Sequence of subunit c of the Na^+-translocating F_1F_O ATPase of *Acetobacterium woodii*: proposal for determinants of Na^+ specificity as revealed by sequence

comparisons. *FEBS Lett* **404,** 269–271 (1997).
55. Lolkema, J. S., Chaban, Y. & Boekema, E. J. Subunit composition, structure, and distribution of bacterial V-type ATPases. *J Bioenerg Biomembr* **35,** 323–335 (2003).


Stomach
Intestine
Fig.S1
caeca
anterior gut
midgut
"*Epulopiscium*"-rich
gut section
posterior gut

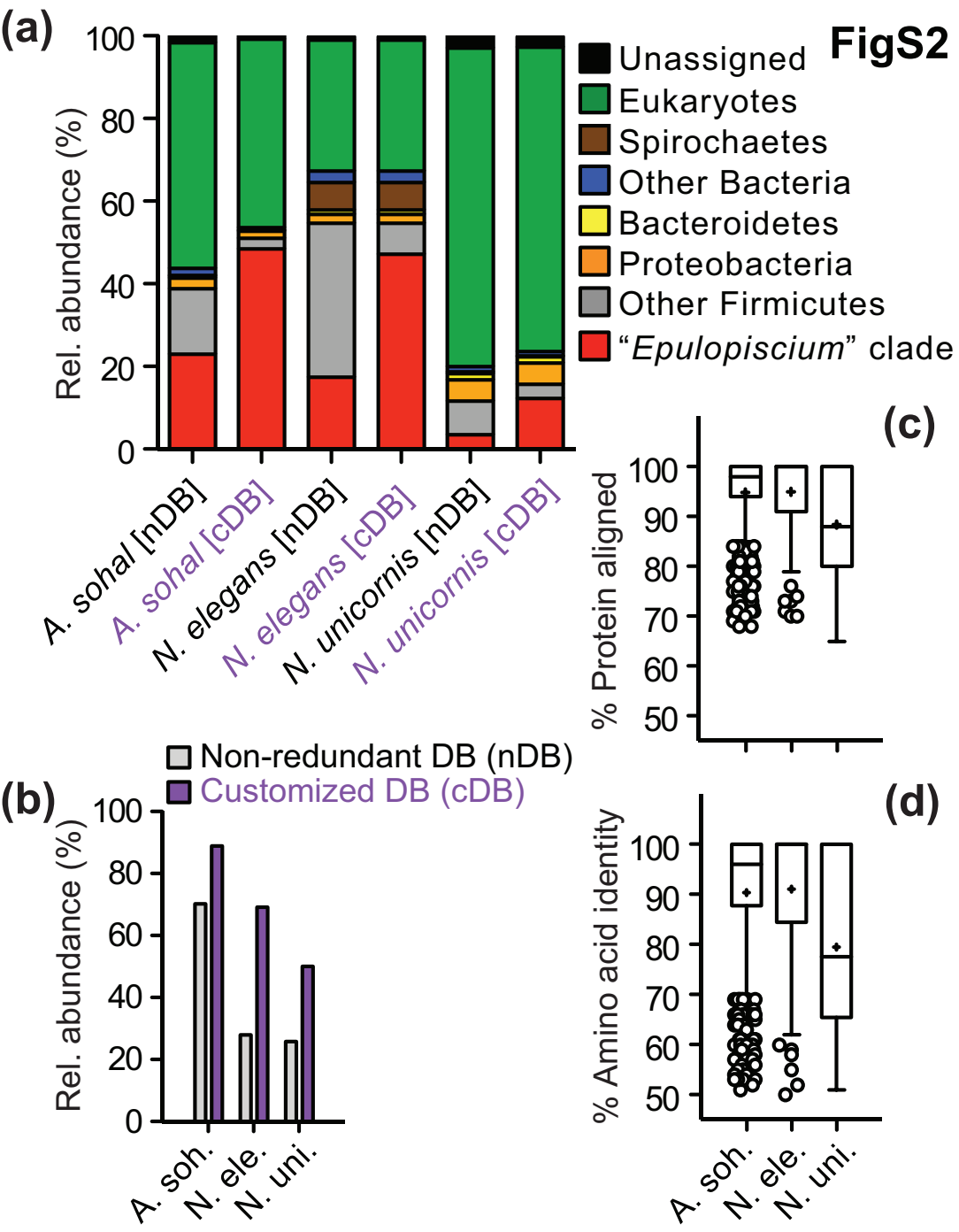

FigS2
(a)
Rel. abundance (%)
100
80
60
40
20
0
Unassigned
Eukaryotes
Spirochaetes
Other Bacteria
Bacteroidetes
Proteobacteria
Other Firmicutes
"Epulopiscium" clade
A. sohal [nDB]
A. sohal [cDB]
N. elegans [nDB]
N. elegans [cDB]
N. unicornis [nDB]
N. unicornis [cDB]
(c)
% Protein aligned
100
90
80
70
60
50
Non-redundant DB (nDB)
Customized DB (cDB)
(b)
Rel. abundance (%)
100
80
60
40
20
0
A. soh.
N. ele.
N. uni.
(d)
% Amino acid identity
100
90
80
70
60
50
A. soh.
N. ele.
N. uni.

Fig. S3

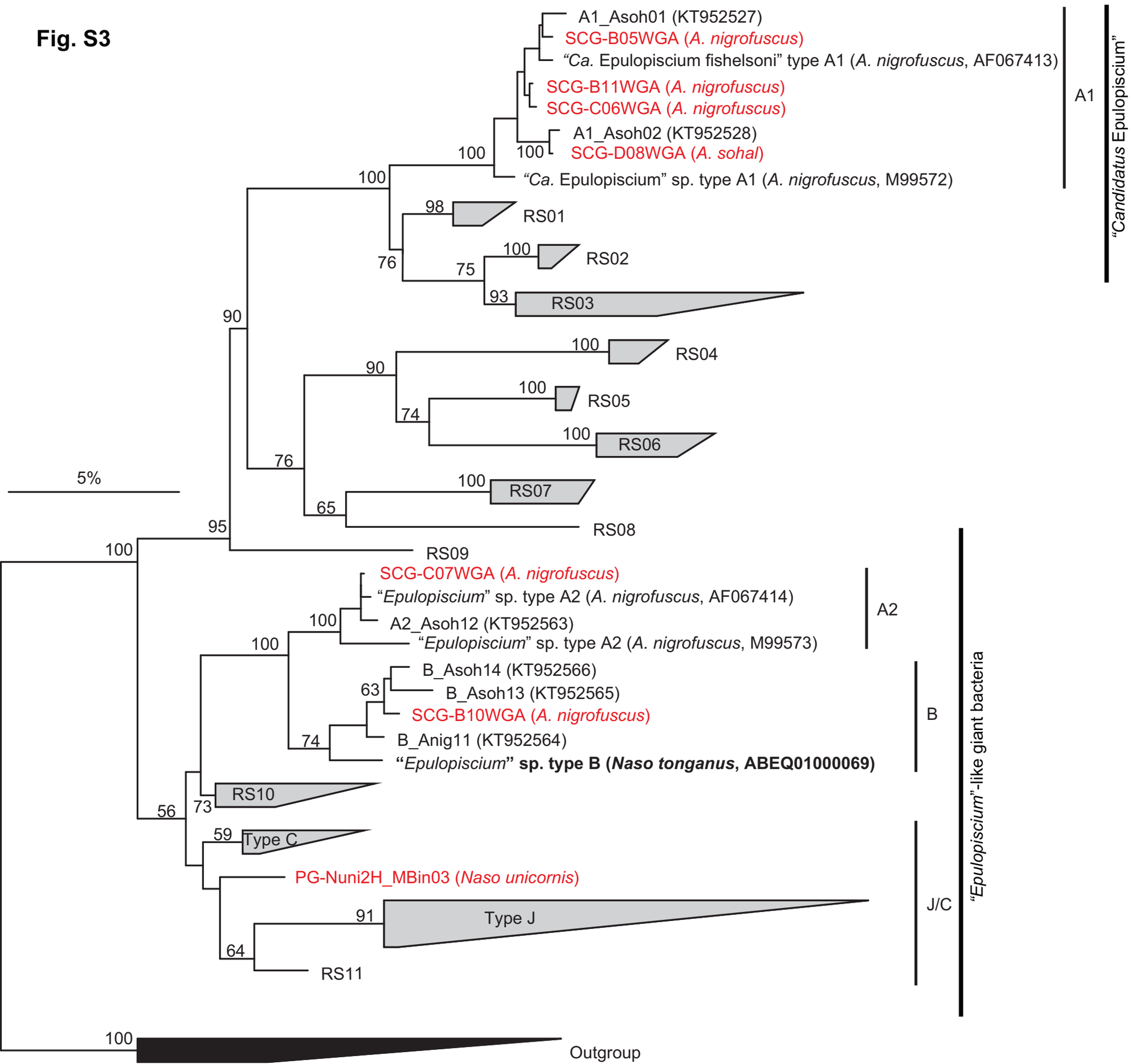
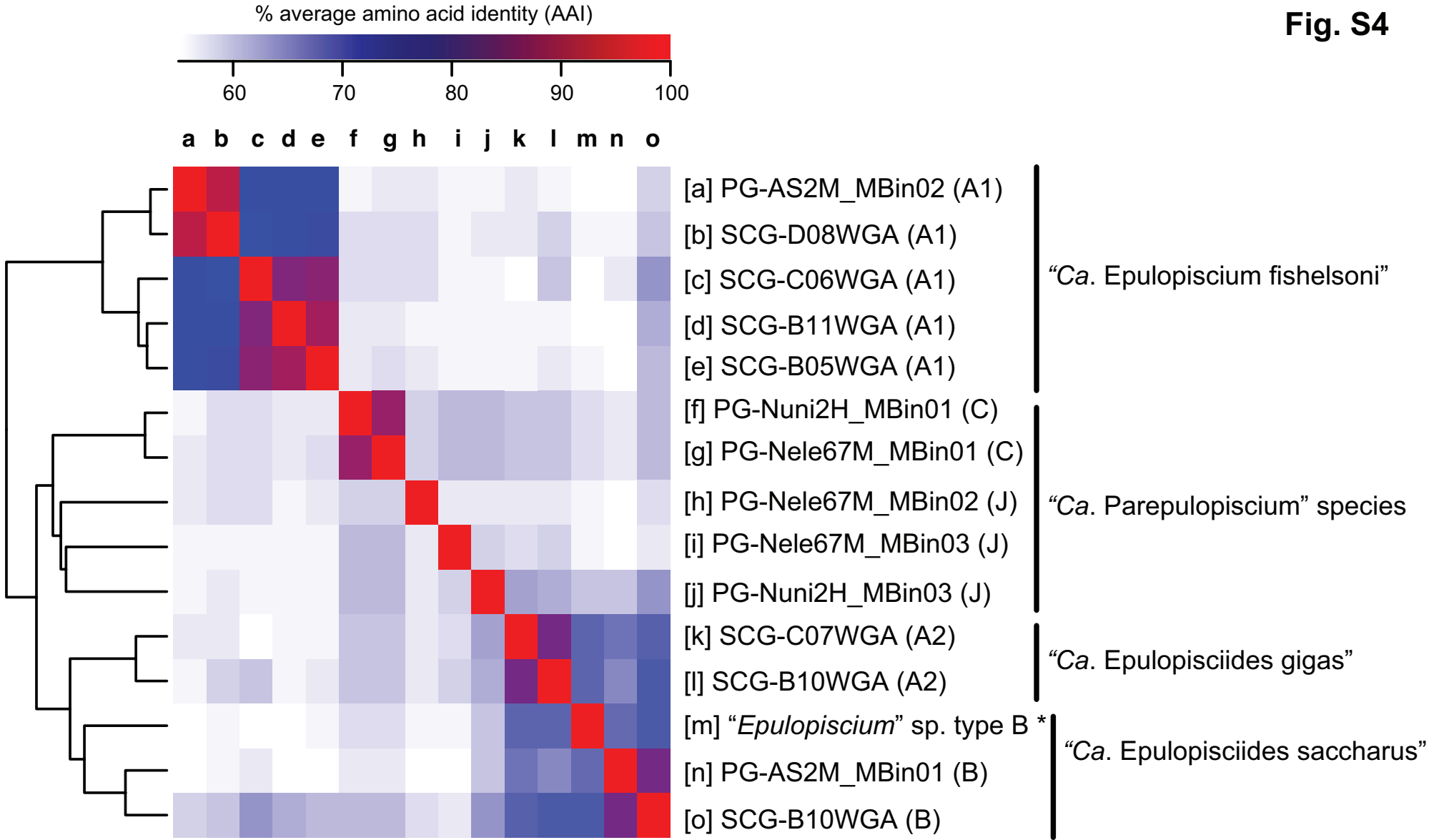
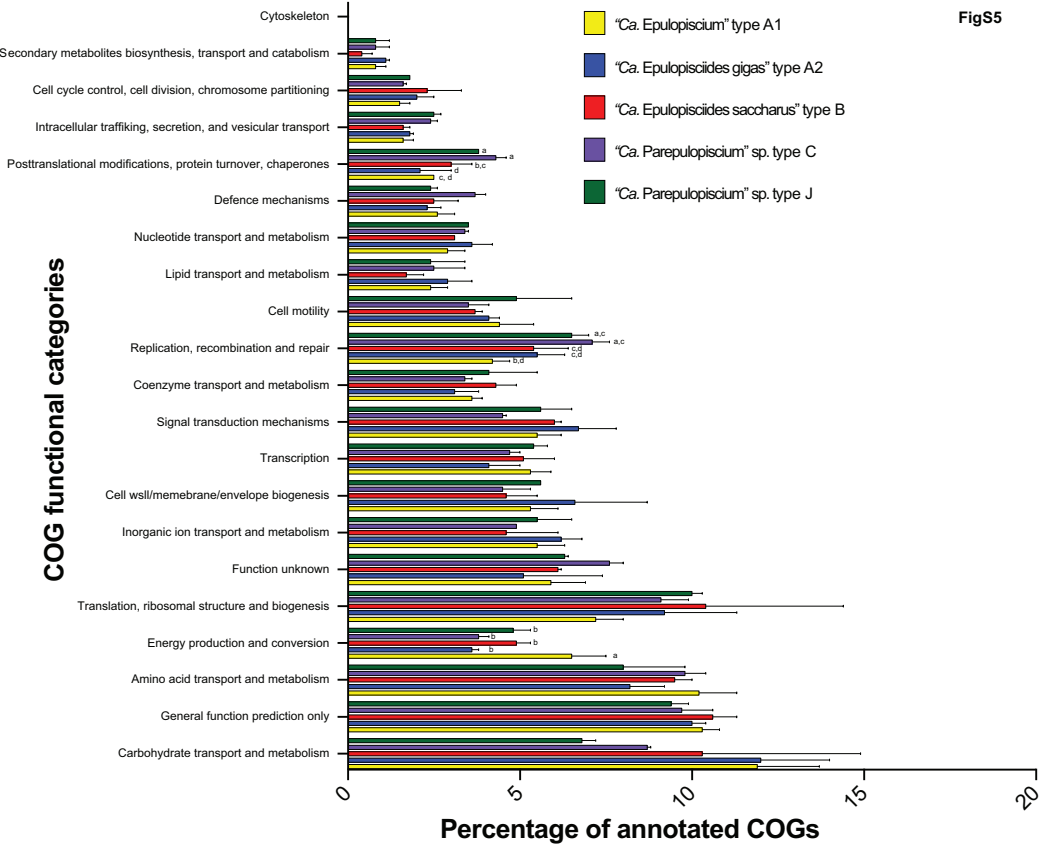
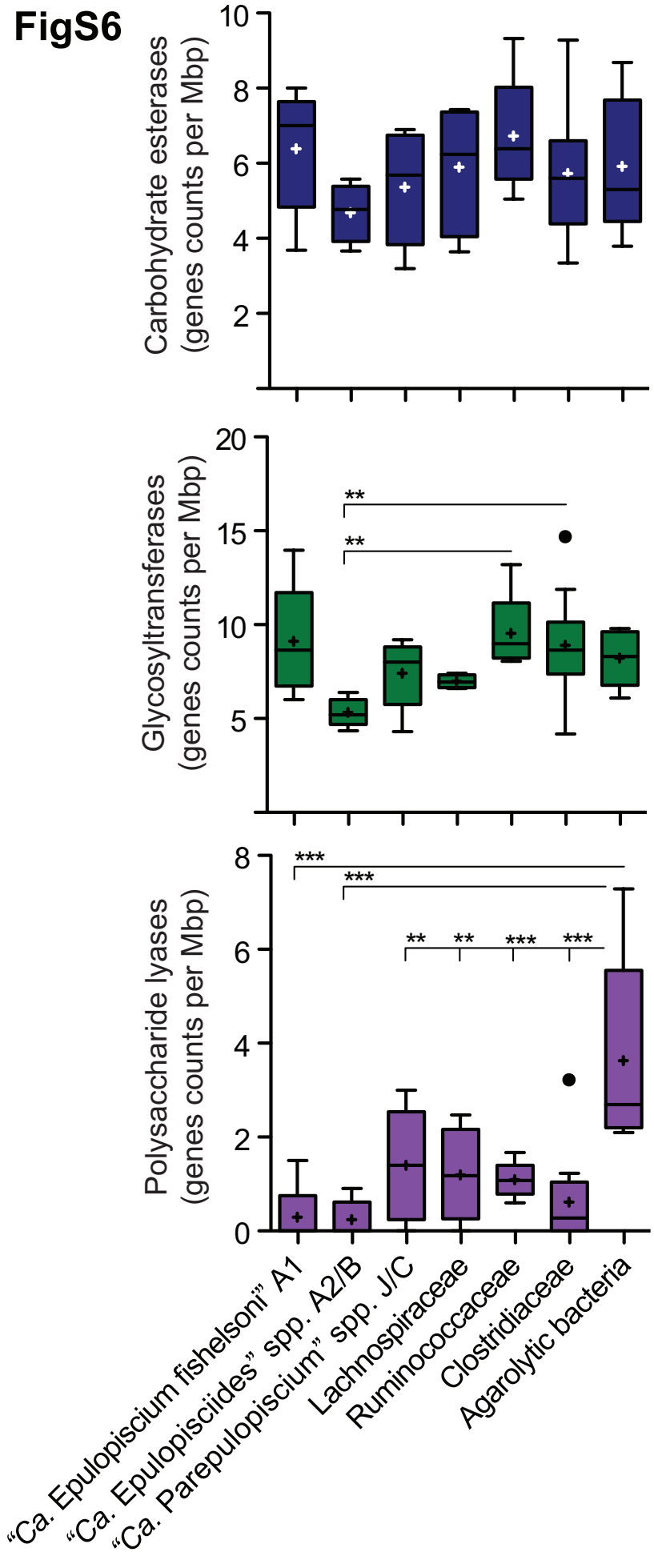

A1_Asoh01 (KT952527)
SCG-B05WGA (A. nigrofuscus)
"Ca. Epulopiscium fishelsoni" type A1 (A. nigrofuscus, AF067413)
SCG-B11WGA (A. nigrofuscus)
SCG-C06WGA (A. nigrofuscus)
A1_Asoh02 (KT952528)
SCG-D08WGA (A. sohal)
"Ca. Epulopiscium" sp. type A1 (A. nigrofuscus, M99572)
A1
"Candidatus Epulopiscium"
RS01
RS02
RS03
RS04
RS05
RS06
RS07
RS08
RS09
5%
SCG-C07WGA (A. nigrofuscus)
"Epulopiscium" sp. type A2 (A. nigrofuscus, AF067414)
A2_Asoh12 (KT952563)
"Epulopiscium" sp. type A2 (A. nigrofuscus, M99573)
A2
B_Asoh14 (KT952566)
B_Asoh13 (KT952565)
SCG-B10WGA (A. nigrofuscus)
B_Anig11 (KT952564)
"Epulopiscium" sp. type B (Naso tonganus, ABEQ01000069)
B
RS10
Type C
PG-Nuni2H_MBin03 (Naso unicornis)
Type J
RS11
J/C
"Epulopiscium"-like giant bacteria
Outgroup

Fig. S4

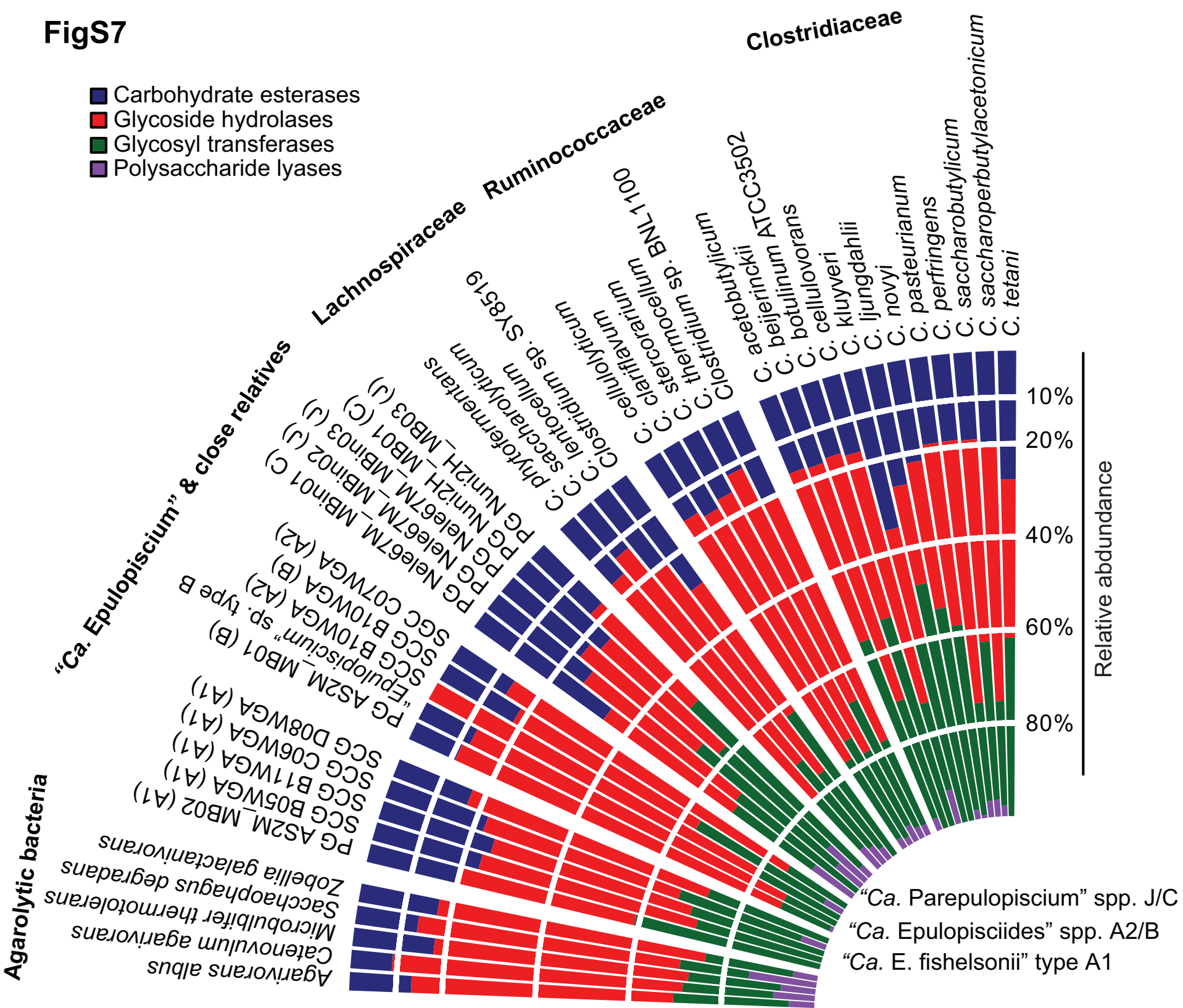

% average amino acid identity (AAI)
60
70
80
90
100
a b c d e f g h i j k l m n o
[a] PG-AS2M_MBin02 (A1)
[b] SCG-D08WGA (A1)
[c] SCG-C06WGA (A1)
[d] SCG-B11WGA (A1)
[e] SCG-B05WGA (A1)
[f] PG-Nuni2H_MBin01 (C)
[g] PG-Nele67M_MBin01 (C)
[h] PG-Nele67M_MBin02 (J)
[i] PG-Nele67M_MBin03 (J)
[j] PG-Nuni2H_MBin03 (J)
[k] SCG-C07WGA (A2)
[l] SCG-B10WGA (A2)
[m] "Epulopiscium" sp. type B *
[n] PG-AS2M_MBin01 (B)
[o] SCG-B10WGA (B)
"Ca. Epulopiscium fishelsoni"
"Ca. Parepulopiscium" species
"Ca. Epulopisciides gigas"
"Ca. Epulopisciides saccharus"

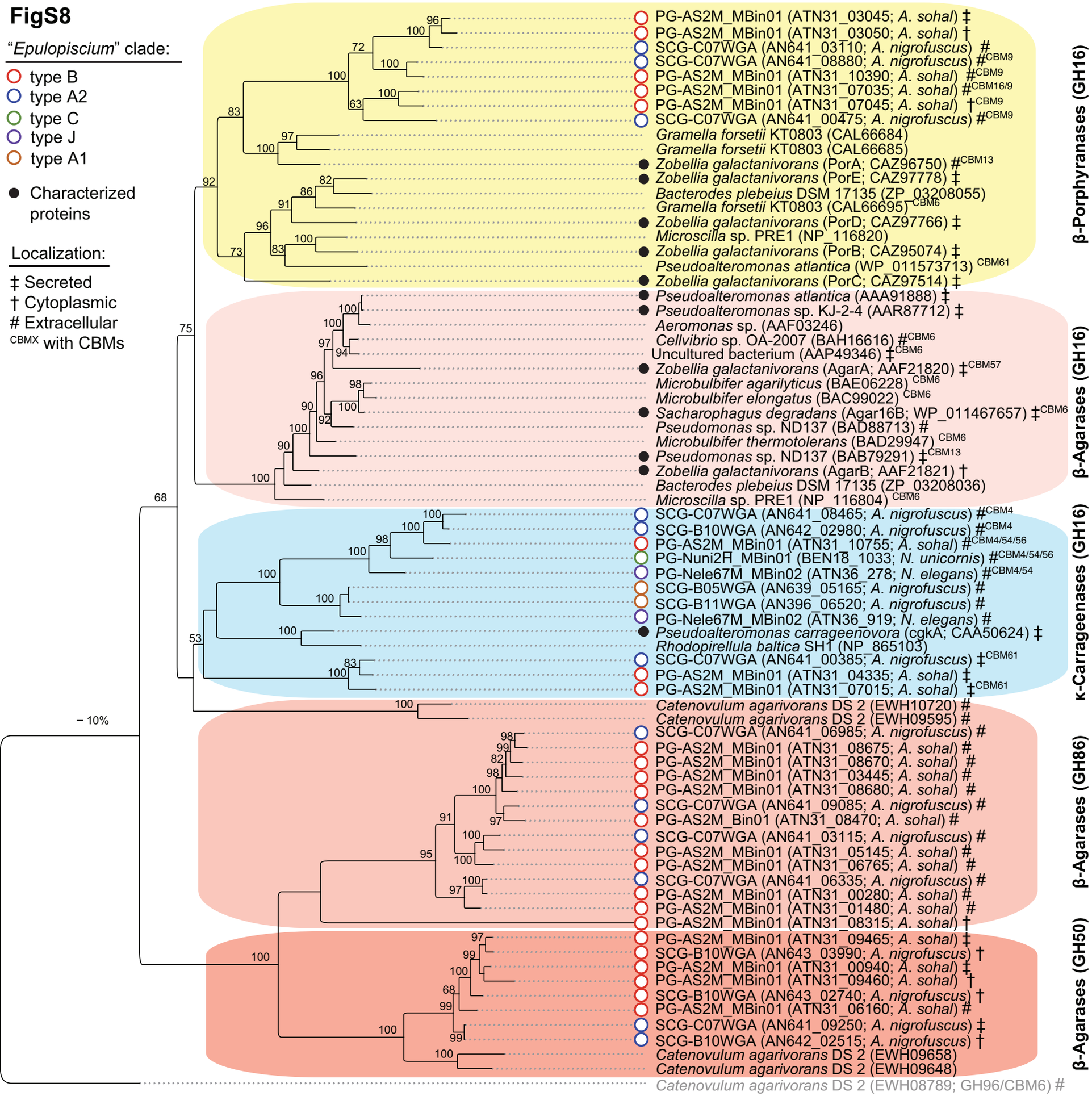

FigS5
"Ca. Epulopiscium" type A1
"Ca. Epulopisciides gigas" type A2
"Ca. Epulopisciides saccharus" type B
"Ca. Parepulopiscium" sp. type C
"Ca. Parepulopiscium" sp. type J
COG functional categories
Cytoskeleton
Secondary metabolites biosynthesis, transport and catabolism
Cell cycle control, cell division, chromosome partitioning
Intracellular traffiking, secretion, and vesicular transport
Posttranslational modifications, protein turnover, chaperones
Defence mechanisms
Nucleotide transport and metabolism
Lipid transport and metabolism
Cell motility
Replication, recombination and repair
Coenzyme transport and metabolism
Signal transduction mechanisms
Transcription
Cell wsll/memebrane/envelope biogenesis
Inorganic ion transport and metabolism
Function unknown
Translation, ribosomal structure and biogenesis
Energy production and conversion
Amino acid transport and metabolism
General function prediction only
Carbohydrate transport and metabolism
0
5
10
15
20
Percentage of annotated COGs

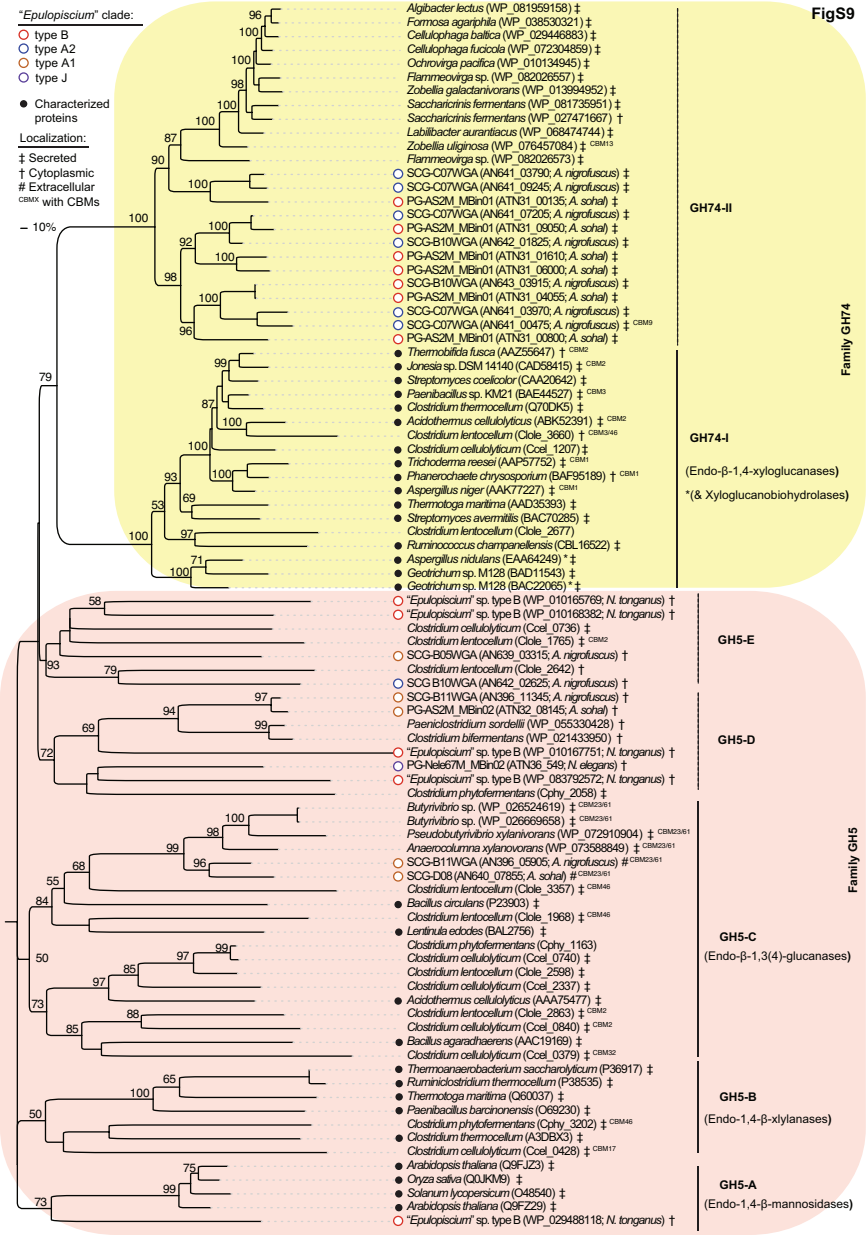
FigS6

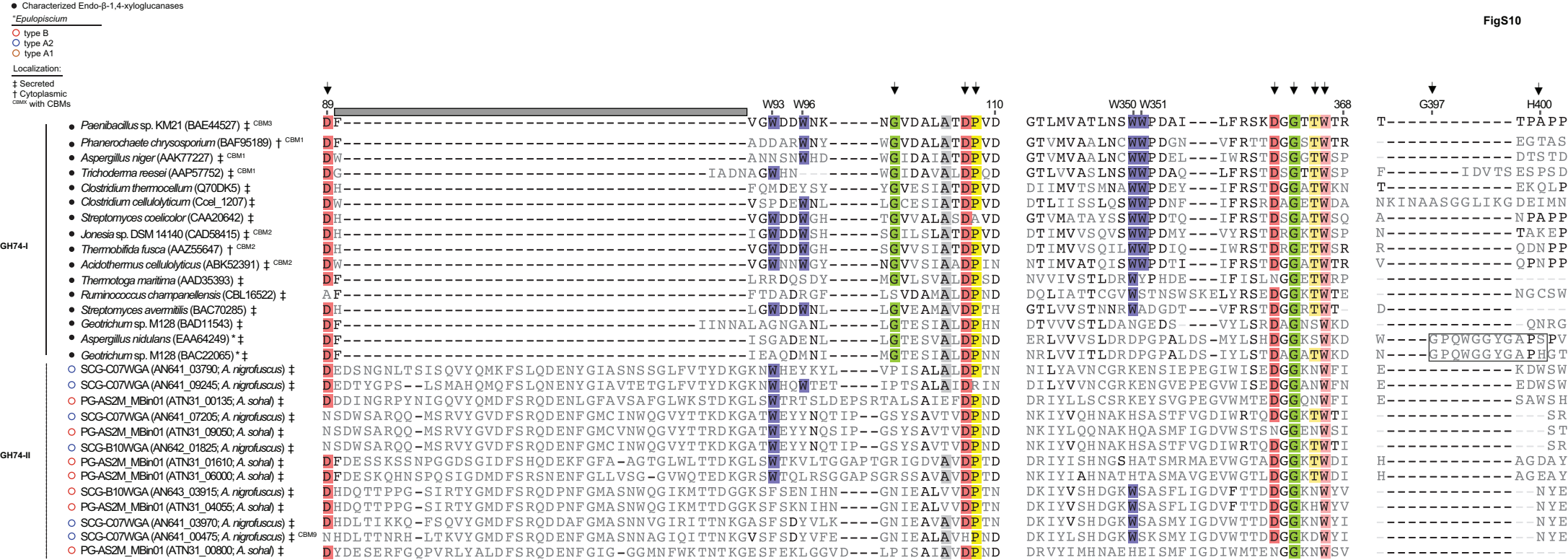

Carbohydrate esterases (genes counts per Mbp)
Glycosyltransferases (genes counts per Mbp)
Polysaccharide lyases (genes counts per Mbp)
**
**
***
***
**
**
***
***
"Ca. Epulopiscium fishelsoni" A1
"Ca. Epulopisciides" spp. A2/B
"Ca. Parepulopiscium" spp. J/C
Lachnospiraceae
Ruminococcaceae
Clostridiaceae
Agarolytic bacteria

FigS7


Carbohydrate esterases
Glycoside hydrolases
Glycosyl transferases
Polysaccharide lyases
Clostridiaceae
Ruminococcaceae
Lachnospiraceae
"Ca. Epulopiscium" & close relatives
Agarolytic bacteria
C. tetani
C. saccharoperbutylacetonicum
C. perfringens
C. pasteurianum
C. novyi
C. ljungdahlii
C. kluyveri
C. cellulovorans
C. botulinum ATCC3502
C. beijerinckii
C. acetobutylicum
Clostridium sp. BNL1100
C. clariflavum
C. cellulolyticum
C. thermocellum
C. sp. SY8519
C. phytofermentans
C. lentocellum
PG Nuni2H_MB03 (J)
PG Nuni2H_MB01 (C)
PG Nele67M_MBin03 (J)
PG Nele67M_MBin02 (J)
PG Nele67M_MBin01 (C)
SCG C07WGA (A2)
SCG B10WGA (B)
SCG B10WGA (A2)
"Epulopiscium" sp. type B
PG AS2M_MB01 (B)
SCG D08WGA (A1)
SCG C06WGA (A1)
SCG B11WGA (A1)
SCG B05WGA (A1)
PG AS2M_MB02 (A1)
Zobellia galactanivorans
Saccharophagus degradans
Microbulbifer thermotolerans
Catenovulum agarivorans
Agarivorans albus
10%
20%
40%
60%
80%
Relative abundance
"Ca. Parepulopiscium" spp. J/C
"Ca. Epulopisciides" spp. A2/B
"Ca. E. fishelsonii" type A1

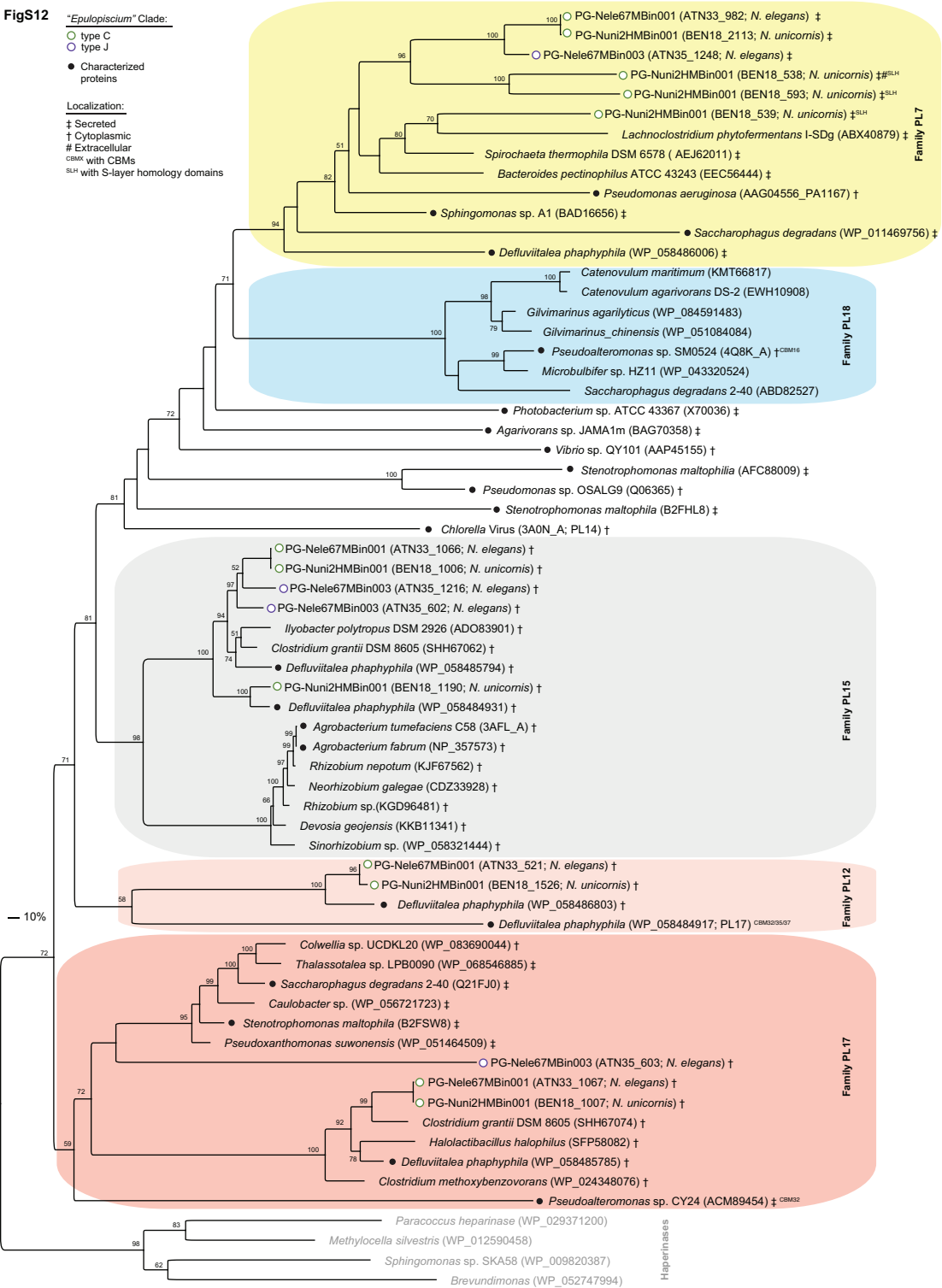

FigS8
"Epulopiscium" clade:
type B
type A2
type C
type J
type A1
Characterized proteins
Localization:
‡ Secreted
† Cytoplasmic
# Extracellular
CBMX with CBMs
PG-AS2M_MBin01 (ATN31_03045; A. sohal) ‡
PG-AS2M_MBin01 (ATN31_03050; A. sohal) †
SCG-C07WGA (AN641_03110; A. nigrofuscus) #
SCG-C07WGA (AN641_08880; A. nigrofuscus) #CBM9
PG-AS2M_MBin01 (ATN31_10390; A. sohal) #CBM9
PG-AS2M_MBin01 (ATN31_07035; A. sohal) #CBM16/9
PG-AS2M_MBin01 (ATN31_07045; A. sohal) †CBM9
SCG-C07WGA (AN641_00475; A. nigrofuscus) #CBM9
Gramella forsetii KT0803 (CAL66684)
Gramella forsetii KT0803 (CAL66685)
Zobellia galactanivorans (PorA; CAZ96750) #CBM13
Zobellia galactanivorans (PorE; CAZ97778) ‡
Bacterodes plebeius DSM 17135 (ZP_03208055)
Gramella forsetii KT0803 (CAL66695) CBM6
Zobellia galactanivorans (PorD; CAZ97766) ‡
Microscilla sp. PRE1 (NP_116820)
Zobellia galactanivorans (PorB; CAZ95074) ‡
Pseudoalteromonas atlantica (WP_011573713) CBM61
Zobellia galactanivorans (PorC; CAZ97514) ‡
β-Porphyranases (GH16)
Pseudoalteromonas atlantica (AAA91888) ‡
Pseudoalteromonas sp. KJ-2-4 (AAR87712) ‡
Aeromonas sp. (AAF03246)
Cellvibrio sp. OA-2007 (BAH16616) #CBM6
Uncultured bacterium (AAP49346) ‡CBM6
Zobellia galactanivorans (AgarA; AAF21820) ‡CBM57
Microbulbifer agarilyticus (BAE06228) CBM6
Microbulbifer elongatus (BAC99022) CBM6
Sacharophagus degradans (Agar16B; WP_011467657) ‡CBM6
Pseudomonas sp. ND137 (BAD88713) #
Microbulbifer thermotolerans (BAD29947) CBM6
Pseudomonas sp. ND137 (BAB79291) ‡CBM13
Zobellia galactanivorans (AgarB; AAF21821) †
Bacterodes plebeius DSM 17135 (ZP_03208036)
Microscilla sp. PRE1 (NP_116804) CBM6
β-Agarases (GH16)
SCG-C07WGA (AN641_08465; A. nigrofuscus) #CBM4
SCG-B10WGA (AN642_02980; A. nigrofuscus) #CBM4
PG-AS2M_MBin01 (ATN31_10755; A. sohal) #CBM4/54/56
PG-Nuni2H_MBin01 (BEN18_1033; N. unicornis) #CBM4/54/56
PG-Nele67M_MBin02 (ATN36_278; N. elegans) #CBM4/54
SCG-B05WGA (AN639_05165; A. nigrofuscus) #
SCG-B11WGA (AN396_06520; A. nigrofuscus) #
PG-Nele67M_MBin02 (ATN36_919; N. elegans) #
Pseudoalteromonas carrageenovora (cgkA; CAA50624) ‡
Rhodopirellula baltica SH1 (NP_865103)
SCG-C07WGA (AN641_00385; A. nigrofuscus) ‡CBM61
PG-AS2M_MBin01 (ATN31_04335; A. sohal) ‡
PG-AS2M_MBin01 (ATN31_07015; A. sohal) ‡CBM61
κ-Carrageenases (GH16)
Catenovulum agarivorans DS 2 (EWH10720) #
Catenovulum agarivorans DS 2 (EWH09595) #
SCG-C07WGA (AN641_06985; A. nigrofuscus) #
PG-AS2M_MBin01 (ATN31_08675; A. sohal) #
PG-AS2M_MBin01 (ATN31_08670; A. sohal) #
PG-AS2M_MBin01 (ATN31_03445; A. sohal) #
PG-AS2M_MBin01 (ATN31_08680; A. sohal) #
SCG-C07WGA (AN641_09085; A. nigrofuscus) #
PG-AS2M_Bin01 (ATN31_08470; A. sohal) #
SCG-C07WGA (AN641_03115; A. nigrofuscus) #
PG-AS2M_MBin01 (ATN31_05145; A. sohal) #
PG-AS2M_MBin01 (ATN31_06765; A. sohal) #
SCG-C07WGA (AN641_06335; A. nigrofuscus) #
PG-AS2M_MBin01 (ATN31_00280; A. sohal) #
PG-AS2M_MBin01 (ATN31_01480; A. sohal) #
PG-AS2M_MBin01 (ATN31_08315; A. sohal) †
β-Agarases (GH86)
PG-AS2M_MBin01 (ATN31_09465; A. sohal) ‡
SCG-B10WGA (AN643_03990; A. nigrofuscus) †
PG-AS2M_MBin01 (ATN31_00940; A. sohal) ‡
PG-AS2M_MBin01 (ATN31_09460; A. sohal) †
SCG-B10WGA (AN643_02740; A. nigrofuscus) †
PG-AS2M_MBin01 (ATN31_06160; A. sohal) #
SCG-C07WGA (AN641_09250; A. nigrofuscus) ‡
SCG-B10WGA (AN642_02515; A. nigrofuscus) †
Catenovulum agarivorans DS 2 (EWH09658)
Catenovulum agarivorans DS 2 (EWH09648)
β-Agarases (GH50)
Catenovulum agarivorans DS 2 (EWH08789; GH96/CBM6) #
– 10%


FigS9
"Epulopiscium" clade:
type B
type A2
type A1
type J
Characterized proteins
Localization:
‡ Secreted
† Cytoplasmic
# Extracellular
CBMX with CBMs
– 10%
Algibacter lectus (WP_081959158) ‡
Formosa agariphila (WP_038530321) ‡
Cellulophaga baltica (WP_029446883) ‡
Cellulophaga fucicola (WP_072304859) ‡
Ochrovirga pacifica (WP_010134945) ‡
Flammeovirga sp. (WP_082026557) ‡
Zobellia galactanivorans (WP_013994952) ‡
Saccharicrinis fermentans (WP_081735951) ‡
Saccharicrinis fermentans (WP_027471667) †
Labilibacter auranticus (WP_068474744) ‡
Zobellia uliginosa (WP_076457084) ‡ CBM13
Flammeovirga sp. (WP_082026573) ‡
SCG-C07WGA (AN641_03790; A. nigrofuscus) ‡
SCG-C07WGA (AN641_09245; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_00135; A. sohai) ‡
SCG-C07WGA (AN641_07205; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_09050; A. sohai) ‡
SCG-B10WGA (AN642_01825; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_01610; A. sohai) ‡
PG-AS2M_MBin01 (ATN31_06000; A. sohai) ‡
SCG-B10WGA (AN643_03915; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_04055; A. sohai) ‡
SCG-C07WGA (AN641_03970; A. nigrofuscus) ‡
SCG-C07WGA (AN641_00475; A. nigrofuscus) ‡ CBM9
PG-AS2M_MBin01 (ATN31_00800; A. sohai) ‡
GH74-II
Thermobifida fusca (AAZ55647) † CBM2
Jonesia sp. DSM 14140 (CAD58415) ‡ CBM2
Streptomyces coelicolor (CAA20642) ‡
Paenibacillus sp. KM21 (BAE44527) ‡ CBM3
Clostridium thermocellum (Q70DK5) ‡
Acidothermus cellulolyticus (ABK52391) ‡ CBM2
Clostridium lentocellum (Clole_3660) † CBM3/46
Clostridium cellulolyticum (Ccel_1207) ‡
Trichoderma reesei (AAP57752) ‡ CBM1
Phanerochaete chrysosporium (BAF95189) † CBM1
Aspergillus niger (AAK77227) ‡ CBM1
Thermotoga maritima (AAD35393) ‡
Streptomyces avermitilis (BAC70285) ‡
Clostridium lentocellum (Clole_2677)
Ruminococcus champanellensis (CBL16522) ‡
Aspergillus nidulans (EAA64249) *‡
Geotrichum sp. M128 (BAD11543) ‡
Geotrichum sp. M128 (BAC22665) *‡
GH74-I
(Endo-β-1,4-xyloglucanases)
*(& Xyloglucanobiohydrolases)
Family GH74
"Epulopiscium" sp. type B (WP_010165769; N. tonganus) †
"Epulopiscium" sp. type B (WP_010168382; N. tonganus) †
Clostridium cellulolyticum (Ccel_0736) ‡
Clostridium lentocellum (Clole_1765) ‡ CBM2
SCG-B05WGA (AN639_03315; A. nigrofuscus) †
Clostridium lentocellum (Clole_2642) †
SCG B10WGA (AN642_02625; A. nigrofuscus) †
GH5-E
SCG-B11WGA (AN396_11345; A. nigrofuscus) †
PG-AS2M_MBin02 (ATN32_08145; A. sohai) †
Paeniclostridium sordellii (WP_055330428) †
Clostridium bifermentans (WP_021433950) †
"Epulopiscium" sp. type B (WP_010167751; N. tonganus) †
PG-Nele67M_MBin02 (ATN36_549; N. elegans) †
"Epulopiscium" sp. type B (WP_083792572; N. tonganus) †
GH5-D
Clostridium phytofermentans (Cphy_2058) ‡
Butyrivibrio sp. (WP_026524619) ‡ CBM23/61
Butyrivibrio sp. (WP_026669658) ‡ CBM23/61
Pseudobutyrivibrio xylanivorans (WP_072910904) ‡ CBM23/61
Anaerocolumna xylanovorans (WP_073588849) ‡ CBM23/61
SCG-B11WGA (AN396_05905; A. nigrofuscus) # CBM23/61
SCG-D08 (AN640_07855; A. sohai) # CBM23/61
Clostridium lentocellum (Clole_3357) ‡ CBM46
Bacillus circulans (P23903) ‡
Clostridium lentocellum (Clole_1968) ‡ CBM46
Lentinula edodes (BAL2756) ‡
Clostridium phytofermentans (Cphy_1163)
Clostridium cellulolyticum (Ccel_0740) ‡
Clostridium lentocellum (Clole_2598) ‡
Clostridium cellulolyticum (Ccel_2337) ‡
Acidothermus cellulolyticus (AAA75477) ‡
Clostridium lentocellum (Clole_2863) ‡ CBM2
Clostridium cellulolyticum (Ccel_0840) ‡ CBM2
Bacillus agaradhaerens (AAC19169) ‡
Clostridium cellulolyticum (Ccel_0379) ‡ CBM32
GH5-C
(Endo-β-1,3(4)-glucanases)
Thermoanaerobacterium saccharolyticum (P36917) ‡
Ruminiclostridium thermocellum (P38535) ‡
Thermotoga maritima (Q60037) ‡
Paenibacillus barcinonensis (O69230) ‡
Clostridium phytofermentans (Cphy_3202) ‡ CBM46
Clostridium thermocellum (A3DBX3) ‡
Clostridium cellulolyticum (Ccel_0428) ‡ CBM17
GH5-B
(Endo-1,4-β-xylanases)
Arabidopsis thaliana (Q9FJZ3) ‡
Oryza sativa (Q0JKM9) ‡
Solanum lycopersicum (O48540) ‡
Arabidopsis thaliana (Q9FZ29) ‡
"Epulopiscium" sp. type B (WP_029488118; N. tonganus) †
GH5-A
(Endo-1,4-β-mannosidases)
Family GH5

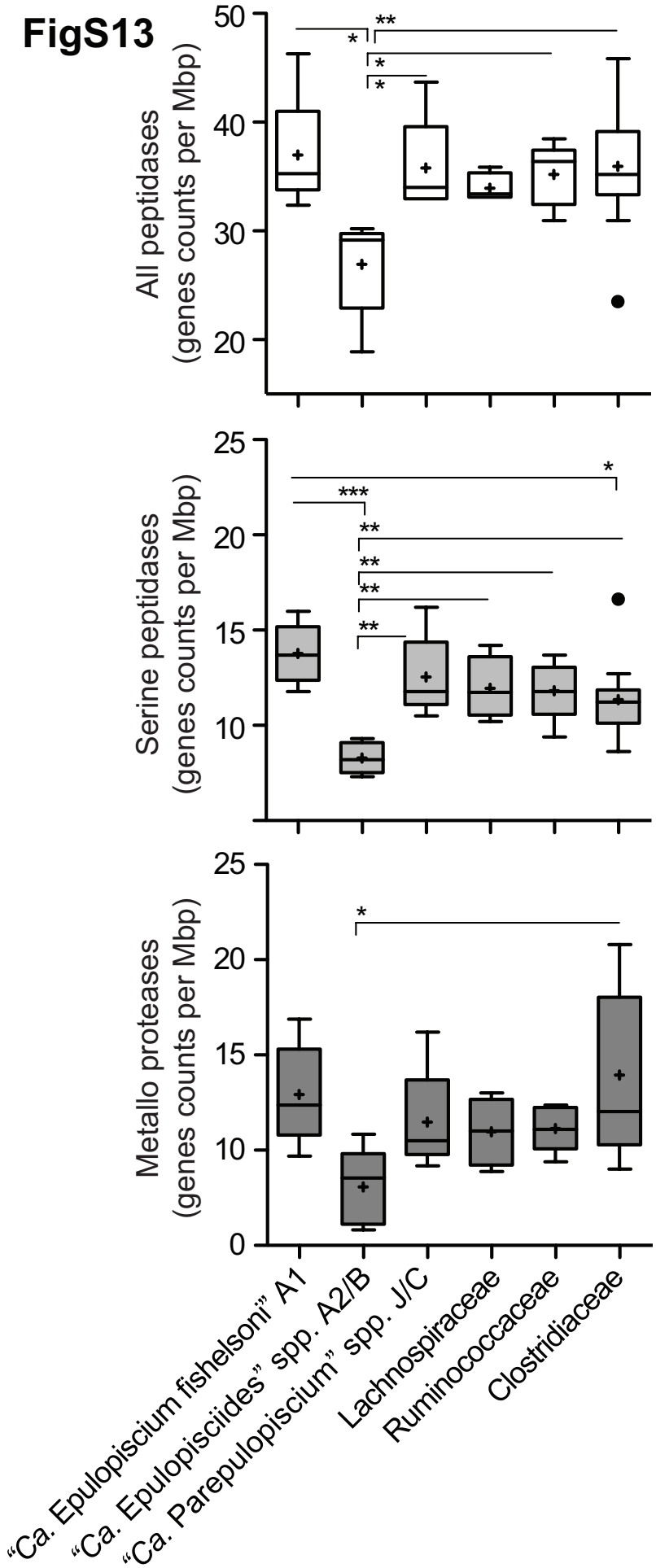

Characterized Endo-β-1,4-xyloglucanases
"Epulopiscium
type B
type A2
type A1
Localization:
‡ Secreted
† Cytoplasmic
CBMX with CBMs
FigS10
89
W93
W96
110
W350
W351
368
G397
H400
GH74-I
GH74-II
Paenibacillus sp. KM21 (BAE44527) ‡ CBM3
Phanerochaete chrysosporium (BAF95189) † CBM1
Aspergillus niger (AAK77227) ‡ CBM1
Trichoderma reesei (AAP57752) ‡ CBM1
Clostridium thermocellum (Q70DK5) ‡
Clostridium cellulolyticum (Ccel_1207) ‡
Streptomyces coelicolor (CAA20642) ‡
Jonesia sp. DSM 14140 (CAD58415) ‡ CBM2
Thermobifida fusca (AAZ55647) † CBM2
Acidothermus cellulolyticus (ABK52391) ‡ CBM2
Thermotoga maritima (AAD35393) ‡
Ruminococcus champanellensis (CBL16522) ‡
Streptomyces avermitilis (BAC70285) ‡
Geotrichum sp. M128 (BAD11543) ‡
Aspergillus nidulans (EAA64249) *‡
Geotrichum sp. M128 (BAC22065) *‡
SCG-C07WGA (AN641_03790; A. nigrofuscus) ‡
SCG-C07WGA (AN641_09245; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_00135; A. sohal) ‡
SCG-C07WGA (AN641_07205; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_09050; A. sohal) ‡
SCG-B10WGA (AN642_01825; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_01610; A. sohal) ‡
PG-AS2M_MBin01 (ATN31_06000; A. sohal) ‡
SCG-B10WGA (AN643_03915; A. nigrofuscus) ‡
PG-AS2M_MBin01 (ATN31_04055; A. sohal) ‡
SCG-C07WGA (AN641_03970; A. nigrofuscus) ‡
SCG-C07WGA (AN641_00475; A. nigrofuscus) ‡ CBM9
PG-AS2M_MBin01 (ATN31_00800; A. sohal) ‡


FigS11
"Epulopiscium" clade:
type B
type A2
Characterized proteins
Localization:
‡ Secreted
† Cytoplasmic
# Extracellular
CBMX with CBMs
PG-AS2M_MBin01 (ATN31_00375; A. sohal) †
PG-AS2M_MBin01 (ATN31_00370; A. sohal) †
SCG-C07WGA (AN641_04775; A. nigrofuscus) †
PG-AS2M_MBin01 (ATN31_05240; A. sohal) †
PG-AS2M_MBin01 (ATN31_11205; A. sohal) †
Clostridium sp. D5 (WP_009003692) †
Ruminococcus sp. SR1\5 (CBL20742) †
Paenibacillus catalpae (SFD49629) †
Clostridium stercorarium (ANX00174) †
Gracilibacillus massiliensis (WP_082631373) †
Clostridium saccharoperbutylacetonicum (WP_015390406) †
Clostridium saccharoperbutylacetonicum (WP_077359864) †
Eisenbergiella tayi (WP_081329914) †
Metagenome from Naso unicornis (Nuni2H_68979_2) †
Clostridium grantii (WP_073337876) †
Caldanaerobius polysaccharolyticus (WP_084295197) †
Xanthomonas campestris (AAM42160) ‡
Caldicellulosiruptor saccharolyticus (WP_011916877) †
Thermoanaerobacter cellulolyticus (WP_045165224) †
Caldicellulosiruptor acetigenus (WP_029229099) †
Granulicella pectinivorans (SFR98348) ‡
SCG-C07WGA (AN641_09750; A. nigrofuscus) †
Cellulophaga algicola (WP_013551231) ‡
Fusarium graminearum (AFR68935) ‡
Family GH29-E
Lilium longiflorum (BAF85832) †
Bacillus halodurans (BAB04561) †
Arabidopsis thaliana (CAB36703) ‡
Bifidobacterium longum (ACJ53393) †
Bacteroides ovatus ATCC 8483 (EDO10805) ‡
Bifidobacterium bifidum (AAQ72464) ‡
Clostridium perfringens ATCC 13124 (ABG82552) ‡
Aspergillus nidulans (EAA59171) ‡
Family GH95
Bifidobacterium longum (ACJ51546) †
Lactobacillus casei BL23 (CAQ67115) †
Thermotoga maritima MSB8 (AAD35394) †
Fusarium oxysporum (AFR68934) ‡
Family GH29-D
Canis lupus familiaris (CAA63362) ‡
Homo sapiens (AAA52481) ‡
Rattus norvegicus (CAA34268) ‡
Drosophila melanogaster (AAM50292) ‡
Dictyostelium discoideum (AAO51149) ‡
Family GH29-C
Clostridium perfringens (ABG83106) ‡ CBM32/51
Arabidopsis thaliana (NP_180377) ‡
Bifidobacterium longum (ACJ53394) † CBM32
Bacteroides thetaiotaomicron (AAO76732) ‡ CBM32
Streptomyces sp. (AAD10477) ‡ CBM32
Bifidobacterium bifidum (BAH80310) ‡ CBM32
Tannerella forsythia (AEW21393) ‡
Lactobacillus casei (CAQ67984) †
Lactobacillus casei (CAQ67877) †
Paenibacillus thiaminolyticus (CBM40947) †
Uncultured bacterium Mfuc5 (AIC77302) †
Uncultured bacterium Mfuc2 (AIC77299) †
Family GH29-B
Bacteroide thetaiotaomicron (AAO78076) ‡
Family GH29-A
Thermotoga maritima (AAD36910; GH94; cellobiose-phosphorylase)
10%

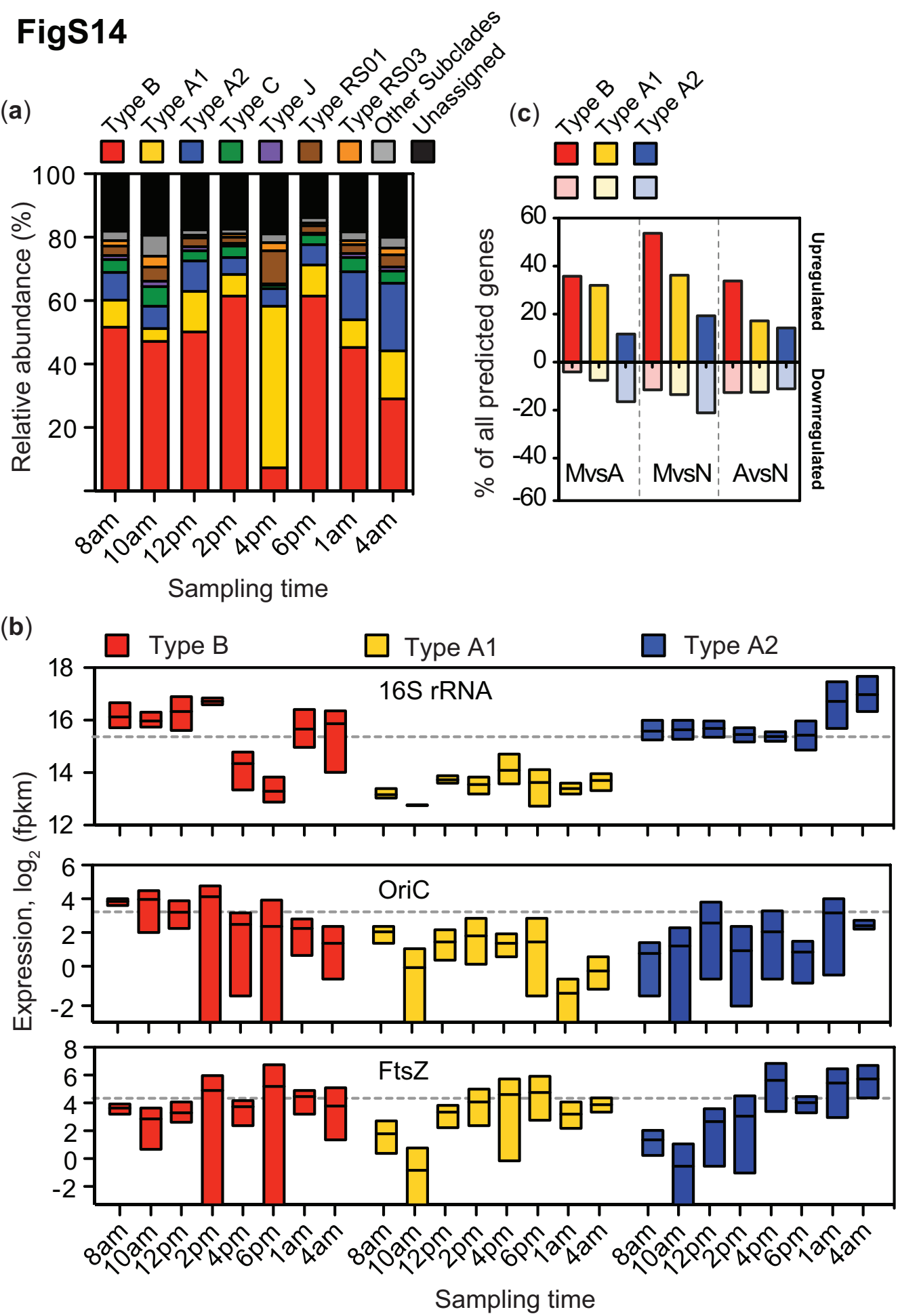
FigS12


"Epulopiscium" Clade:
type C
type J
Characterized proteins
Localization:
‡ Secreted
† Cytoplasmic
# Extracellular
CBMX with CBMs
SLH with S-layer homology domains
PG-Nele67MBin001 (ATN33_982; N. elegans) ‡
PG-Nuni2HMBin001 (BEN18_2113; N. unicornis) ‡
PG-Nele67MBin003 (ATN35_1248; N. elegans) ‡
PG-Nuni2HMBin001 (BEN18_538; N. unicornis) ‡#SLH
PG-Nuni2HMBin001 (BEN18_593; N. unicornis) ‡SLH
PG-Nuni2HMBin001 (BEN18_539; N. unicornis) ‡SLH
Lachnoclostridium phytofermentans I-SDg (ABX40879) ‡
Spirochaeta thermophila DSM 6578 (AEJ62011) ‡
Bacteroides pectinophilus ATCC 43243 (EEC56444) ‡
Pseudomonas aeruginosa (AAG04556_PA1167) †
Sphingomonas sp. A1 (BAD16656) ‡
Saccharophagus degradans (WP_011469756) ‡
Defluviitalea phaphyphila (WP_058486006) ‡
Family PL7
Catenovulum maritimum (KMT66817)
Catenovulum agarivorans DS-2 (EWH10908)
Gilvimarinus agarilyticus (WP_084591483)
Gilvimarinus_chinensis (WP_051084084)
Pseudoalteromonas sp. SM0524 (4Q8K_A) †CBM16
Microbulbifer sp. HZ11 (WP_043320524)
Saccharophagus degradans 2-40 (ABD82527)
Family PL18
Photobacterium sp. ATCC 43367 (X70036) ‡
Agarivorans sp. JAMA1m (BAG70358) ‡
Vibrio sp. QY101 (AAP45155) †
Stenotrophomonas maltophilia (AFC88009) ‡
Pseudomonas sp. OSALG9 (Q06365) †
Stenotrophomonas maltophila (B2FHL8) ‡
Chlorella Virus (3A0N_A; PL14) †
PG-Nele67MBin001 (ATN33_1066; N. elegans) †
PG-Nuni2HMBin001 (BEN18_1006; N. unicornis) †
PG-Nele67MBin003 (ATN35_1216; N. elegans) †
PG-Nele67MBin003 (ATN35_602; N. elegans) †
Ilyobacter polytropus DSM 2926 (ADO83901) †
Clostridium grantii DSM 8605 (SHH67062) †
Defluviitalea phaphyphila (WP_058485794) †
PG-Nuni2HMBin001 (BEN18_1190; N. unicornis) †
Defluviitalea phaphyphila (WP_058484931) †
Agrobacterium tumefaciens C58 (3AFL_A) †
Agrobacterium fabrum (NP_357573) †
Rhizobium nepotum (KJF67562) †
Neorhizobium galegae (CDZ33928) †
Rhizobium sp.(KGD96481) †
Devosia geojensis (KKB11341) †
Sinorhizobium sp. (WP_058321444) †
Family PL15
PG-Nele67MBin001 (ATN33_521; N. elegans) †
PG-Nuni2HMBin001 (BEN18_1526; N. unicornis) †
Defluviitalea phaphyphila (WP_058486803) †
Defluviitalea phaphyphila (WP_058484917; PL17) CBM32/35/37
Family PL12
10%
Colwellia sp. UCDKL20 (WP_083690044) †
Thalassotalea sp. LPB0090 (WP_068546885) ‡
Saccharophagus degradans 2-40 (Q21FJ0) ‡
Caulobacter sp. (WP_056721723) ‡
Stenotrophomonas maltophila (B2FSW8) ‡
Pseudoxanthomonas suwonensis (WP_051464509) ‡
PG-Nele67MBin003 (ATN35_603; N. elegans) †
PG-Nele67MBin001 (ATN33_1067; N. elegans) †
PG-Nuni2HMBin001 (BEN18_1007; N. unicornis) †
Clostridium grantii DSM 8605 (SHH67074) †
Halolactibacillus halophilus (SFP58082) †
Defluviitalea phaphyphila (WP_058485785) †
Clostridium methoxybenzovorans (WP_024348076) †
Pseudoalteromonas sp. CY24 (ACM89454) ‡ CBM32
Family PL17
Paracoccus heparinase (WP_029371200)
Methylocella silvestris (WP_012590458)
Sphingomonas sp. SKA58 (WP_009820387)
Brevundimonas (WP_052747994)
Haperinases

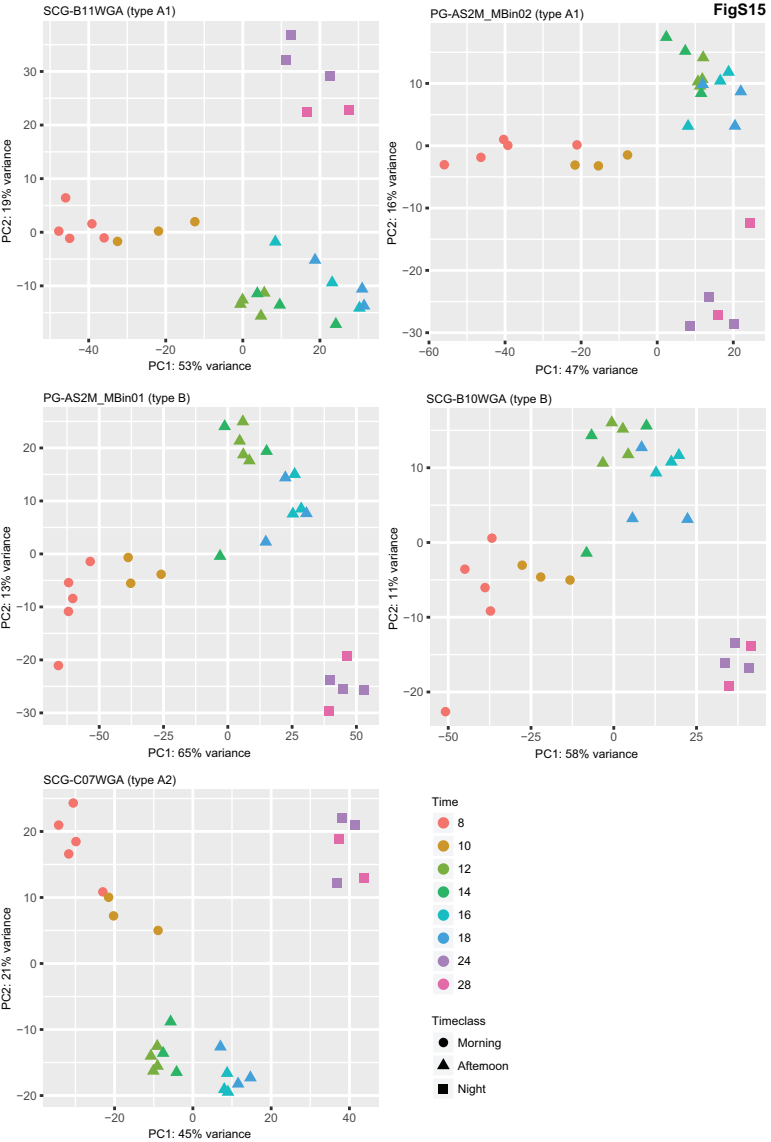
FigS13


All peptidases
(genes counts per Mbp)
50
40
30
20
*
**
*
*
Serine peptidases
(genes counts per Mbp)
25
20
15
10
*
***
**
**
**
**
Metallo proteases
(genes counts per Mbp)
25
20
15
10
0
*
"Ca. Epulopiscium fishelsoni" A1
"Ca. Epulopisciides" spp. A2/B
"Ca. Parepulopiscium" spp. J/C
Lachnospiraceae
Ruminococcaceae
Clostridiaceae

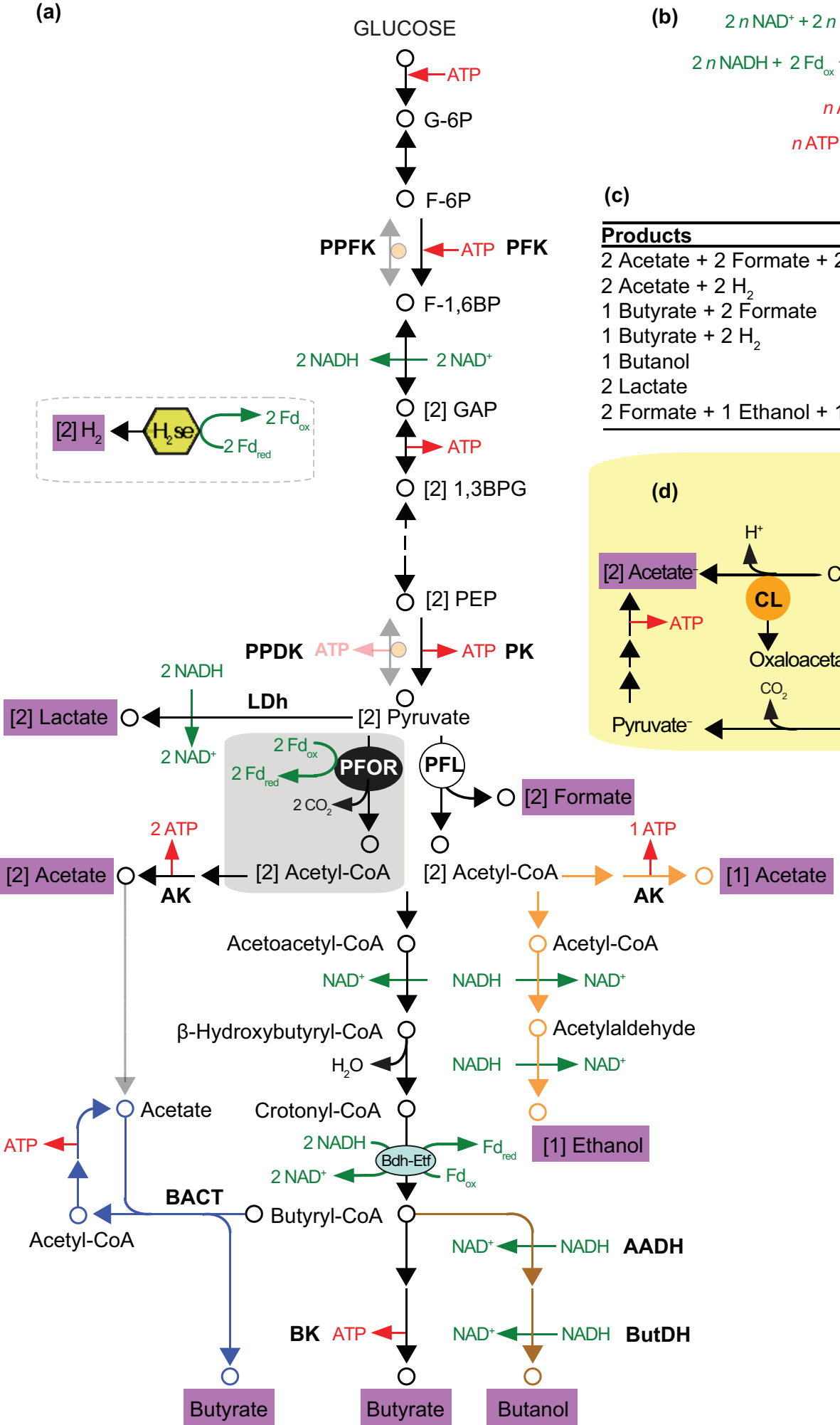
FigS14

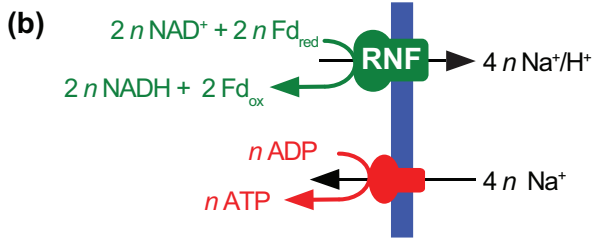

(a)
Type B
Type A1
Type A2
Type C
Type J
Type RS01
Type RS03
Other Subclades
Unassigned
Relative abundance (%)
100
80
60
40
20
8am
10am
12pm
2pm
4pm
6pm
1am
4am
Sampling time
(c)
Type B
Type A1
Type A2
% of all predicted genes
60
40
20
0
-20
-40
-60
MvsA
MvsN
AvsN
Upregulated
Downregulated
(b)
Type B
Type A1
Type A2
16S rRNA
OriC
FtsZ
Expression, log2 (fpkm)
18
16
14
12
6
4
2
0
-2
8
6
4
2
0
-2
8am
10am
12pm
2pm
4pm
6pm
1am
4am
Sampling time

FigS15


SCG-B11WGA (type A1)
PC2: 19% variance
PC1: 53% variance
PG-AS2M_MBin02 (type A1)
PC2: 16% variance
PC1: 47% variance
PG-AS2M_MBin01 (type B)
PC2: 13% variance
PC1: 65% variance
SCG-B10WGA (type B)
PC2: 11% variance
PC1: 58% variance
SCG-C07WGA (type A2)
PC2: 21% variance
PC1: 45% variance
Time
8
10
12
14
16
18
24
28
Timeclass
Morning
Afternoon
Night

FigS16

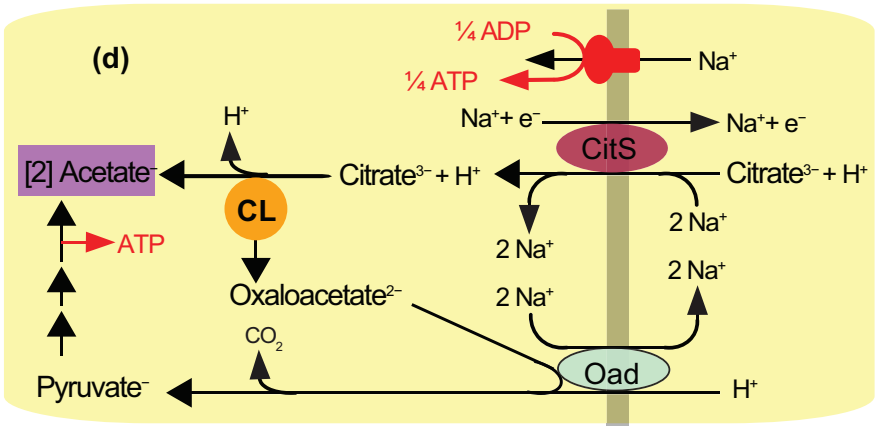




(c)

ATP yield/glucose

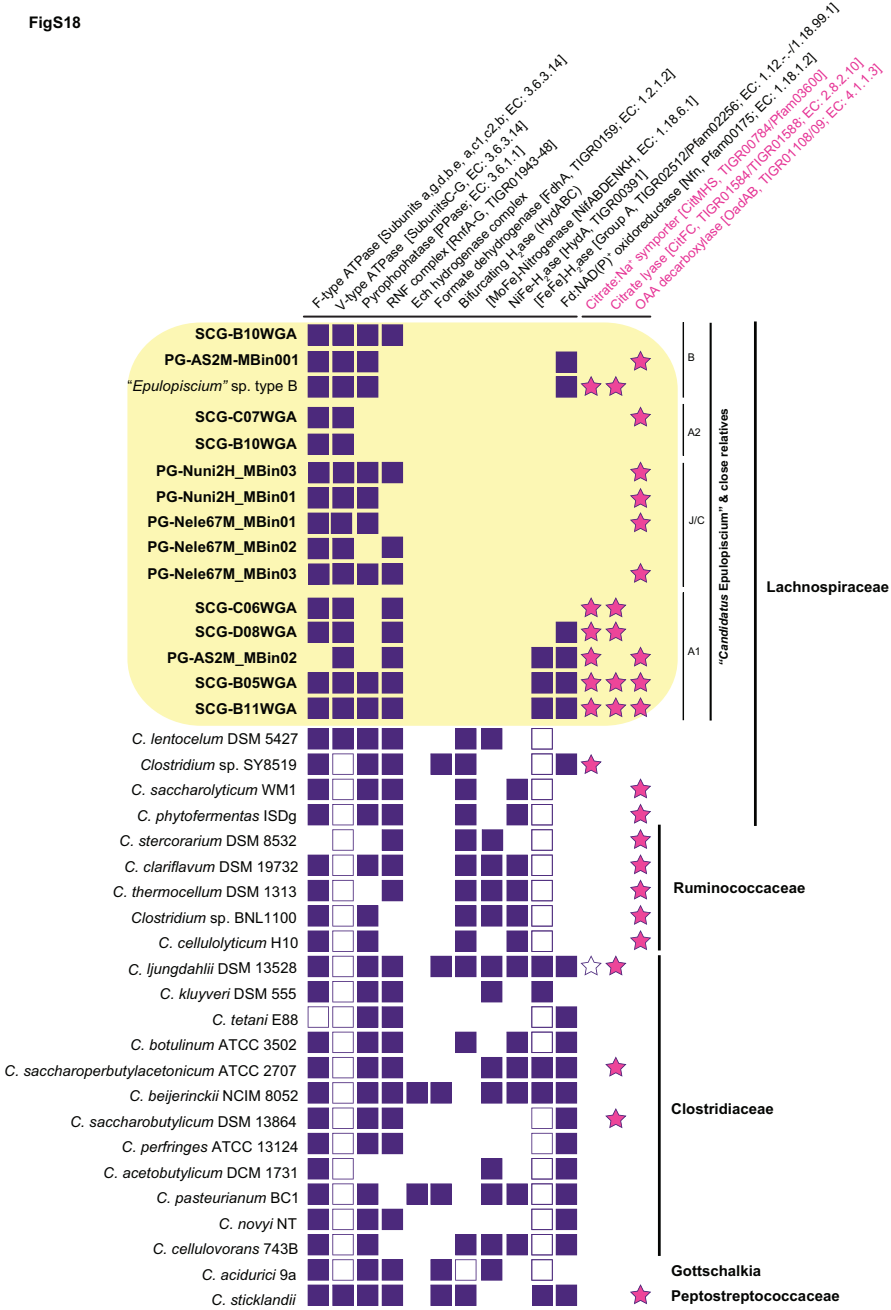
Products	ETP	SLP	Total
2 Acetate + 2 Formate + 2 H_2	–1.0	4	3.0
2 Acetate + 2 H_2	–1.0	4	3.0
1 Butyrate + 2 Formate	0.5	3	3.5
1 Butyrate + 2 H_2	0.5	3	3.5
1 Butanol	1.5	2	3.5
2 Lactate	0	2	2.0
2 Formate + 1 Ethanol + 1 Acetate	0	3	3.0



FigS17


SCG_B05WGA (AN639_02505) -METQIDGKALILACSAIGSGLAMIAGIGPGIGQGYAAGKGAEAVGRQPEAQSDVVRTML 59
SCG_B11WGA (AN396_13075) -METQIDGKALILACSAIGSGLAMIAGIGPGIGQGYAAGKGAEAVGRQPEAQSDVVRTML 59
SCG_D08WGA (AN640_07445) -METQIDGKALILACSAIGSGLAMIAGIGPGIGQGYAAGKGAEAVGRQPEAQSDVVRTML 59
Nele67M_MBin02 (ATN36_00400) MEINAIDGKALILACSAIGAGLAMISGIGPGIGQGYAAGKGAEGVGRQPEAQSDIVRTML 60
Acetobacterium woodii (U10505) -----MEGLDFIKACSAIGAGIAMIAGVGPGIGQGFAAGKGAEAVGRQPEAQSDIIRTML 55
Propionigenium modestum (X53845) --MDMVLAKTVVLAASAVGAGAAMIAGIGPGVGQGYAAGKAVESVARQPEAKGDIISTMV 58
Escherichia coli (M25464) --MENL-NMDLLYMAAAVM---MGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFF 54
SCG_B05WGA (AN639_02505) LGAAVAETTGIYGLIVAIILLFANPLITKYMEIM 93
SCG_B11WGA (AN396_13075) LGAAVAETTGIYGLIVAIILLFANPLITKYMEIM 93
SCG_D08WGA (AN640_07445) LGAAVAETTGIYGLIVAIILLFANPLITKYMDIM 93
Nele67M_MBin02 (ATN36_00400) LGAAVAETTGIYGLIVAIILLFANPLISTYMGLM 94
Acetobacterium woodii (U10505) LGAAVAETTGIYGLIVALILLFANPFF------- 82
Propionigenium modestum (X53845) LGQAIAESTGIYSLVIALILLYANPFVGLLG--- 89
Escherichia coli (M25464) IVMGLVDAIPMIAVGLGLYVMFAVA--------- 79

FigS18


F-type ATPase [Subunits a,g,d,b,e, a,c1,c2,b; EC: 3.6.3.14]
V-type ATPase [SubunitsC-G; EC: 3.6.3.14]
Pyrophosphatase [PPase; EC: 3.6.1.1]
RNF complex [RnfA-G, TIGR01943-48]
Ech hydrogenase complex
Formate dehydrogenase [FdhA, TIGR0159; EC: 1.2.1.2]
Bifurcating H2ase (HydABC)
[MoFe]-Nitrogenase [NifABDENKH, EC: 1.18.6.1]
NiFe-H2ase [HydA, TIGR00391]
[FeFe]-H2ase [Group A, TIGR02512/Pfam02256; EC: 1.12.-.-/1.18.99.1]
Fd.NAD(P)+ oxidoreductase [Nfn, Pfam00175; EC: 1.18.1.2]
Citrate/Na+ symporter [CitMHS, TIGR00784/Pfam03600]
Citrate lyase [CitEFC, TIGR01584/TIGR01588; EC: 2.8.2.10]
OAA decarboxylase [OadAB, TIGR01108/09; EC: 4.1.1.3]
SCG-B10WGA
PG-AS2M-MBin001
"Epulopiscium" sp. type B
SCG-C07WGA
SCG-B10WGA
PG-Nuni2H_MBin03
PG-Nuni2H_MBin01
PG-Nele67M_MBin01
PG-Nele67M_MBin02
PG-Nele67M_MBin03
SCG-C06WGA
SCG-D08WGA
PG-AS2M_MBin02
SCG-B05WGA
SCG-B11WGA
B
A2
J/C
A1
"Candidatus Epulopiscium" & close relatives
Lachnospiraceae
C. lentocelum DSM 5427
Clostridium sp. SY8519
C. saccharolyticum WM1
C. phytofermentas ISDg
C. stercorarium DSM 8532
C. clariflavum DSM 19732
C. thermocellum DSM 1313
Clostridium sp. BNL1100
C. cellulolyticum H10
Ruminococcaceae
C. ljungdahlii DSM 13528
C. kluyveri DSM 555
C. tetani E88
C. botulinum ATCC 3502
C. saccharoperbutylacetonicum ATCC 2707
C. beijerinckii NCIM 8052
C. saccharobutylicum DSM 13864
C. perfringes ATCC 13124
C. acetobutylicum DCM 1731
C. pasteurianum BC1
C. novyi NT
C. cellulovorans 743B
Clostridiaceae
C. acidurici 9a
C. sticklandii
Gottschalkia
Peptostreptococcaceae