Supporting Information

Pope et al. 10.1073/pnas.1005297107

SI Materials and Methods

Cell Dissociation and DNA Extraction. Before cell dissociation and DNA extraction, a subsample of each digesta sample was pooled and hereafter is referred to as T1 (November 2006) and T2 (May 2007). To desorb and recover those microbes adherent to plant biomass, 5 to 10 g of the pooled samples was centrifuged at 12,000 × *g* for 2 min, and the pellet was resuspended in 15 mL dissociation buffer [0.1% Tween 80, 1% methanol and 1% tertiary butanol (vol/vol), pH 2] (1). This mixture was vortexed for 30 s and centrifuged at low speed for 20 s to sediment the large plant particles; this process was repeated two to three more times, with the supernatants collected and pooled. Microbial biomass was collected by centrifugation at 12,000 × *g* for 5 min and the cell pellets were resuspended in 1 mL of sterile 10 mM Tris-HC1 (pH 8.0) 1 M NaCl, then subjected to one final low speed centrifugation for 20 s to remove any residual particulate matter. The cells were finally harvested by centrifugation at 12,000 × *g* for 5 min.

The cell pellets (~200 mg wet-weight) were resuspended in 700 μL TE buffer and incubated at 75 °C for 10 min to inactivate nucleases. Cell lysis was performed by adding lysozyme (1 mg/mL)/mutanolysin (20 U) and achromopeptidase (1 mg/mL) to these cell suspensions and incubation at 37 °C for 90 min. Next, SDS was added to give a final concentration of 1% (wt/vol), 0.20 mg proteinase K was also added, and the mixture was incubated at 55 °C for 90 min. Next, NaCl and CTAB were added to give final concentrations of 0.7 M and 2% (wt/vol) respectively, and the mixture was incubated at 70 °C for 10 min. Following phenol: chloroform:isoamylalcohol and chloroform extractions, the DNA was precipitated with two volumes of 95% ethanol, washed with 70% ethanol, and the pellet air-dried and resuspended in TE buffer (pH 8.0) at a final concentration ~0.5 μg/μL.

16S rRNA Gene PCR Clone Libraries. Two *rrs* clone libraries were prepared from the pooled metagenomic DNA samples by using two different primer pairs broadly targeting the bacterial domain: 27F (5′-AGA GTT TGA TCC TGG CTC AG-3′) and 1492R (5′-GGT TAC CTT GTT ACG ACT T-3′); and GM3 (5′-AGA GTT TGA TCM TGG C-3′) and GM4 (5′-TAC CTT GTT ACG ACT T-3′) (2). The PCR amplicons were cloned into the vector pCR4-TOPO (Invitrogen Corp.) and plated onto Carbenicillin-containing (150 μg/mL) LB agar plates, and two 384-well microtitre plates were picked. The propagated clones were end-sequenced using vector-based primers, and the sequence reads were trimmed, assembled, and quality-checked using the genelib software package (E. Kirton; Joint Genome Institute, Walnut Creek, CA). Thirty-one putative chimeras were identified using Bellerophon (3) and Chimera Check (4) and excluded from the dataset. A total of 663 near-complete bacterial *rrs* gene sequences passed the quality and chimera filters and were used in the subsequent analyses.

Metagenome Processing: Shotgun Library Preparation, Sequencing, and Assembly. Shotgun libraries from the Tammar genomic DNA were prepared from each of the pooled samples T1 and T2: a 2- to 4-kb insert library cloned into pUC18 and a roughly 36-kb insert fosmid library cloned in pCC1Fos (Epicentre Corp.). Libraries were sequenced with BigDye Terminators v3.1 and resolved with ABI PRISM 3730 (ABI) sequencers. A total of 121,728 reads (60,672 from T1 and 61,056 from T2) comprising 87.12 megabases (Mb) of phred Q20 sequence were generated from the small insert library. Sequence reads from T1 and T2 were trimmed with LUCY v. 1.19p (5), resulting in a total of 106,913 reads comprising 82.7 Mb, then pooled together and assembled with the Paracel Genome Assembler (PGA version 2.62, www.paracel.com). The resulting pooled assembly consisted of 12,664 contigs, of which the longest was 27.9 kbp long and contained 237 reads (average read depth 7.9×). Approximately 38% of the reads remained as singlets.

Full Fosmid Sequencing and Assembly. Based on a number of functional and hybridization-based screens, 98 fosmids were chosen for sequencing. The individual fosmids were induced to increase their copy number following Epicentre protocols, and the fosmid DNA purified using Qiagen MiniPrep columns. Equimol amounts of the fosmids were pooled together (~20 μg total DNA) and both a 3-kb paired-end library and a 454 standard shotgun library were constructed. Both libraries were directly sequenced with the 454 Life Sciences Genome Sequencer GS FLX and the libraries produced ~700 Mbp of data with an average read length of 375 bp. Duplicate removal and splitting of paired reads reduced the dataset to 560 Mbp in 2,077,631 reads. The Newbler assembly tool was applied to these data and 33 of the fosmid inserts were completely assembled, another 39 fosmid inserts were reconstructed from two or more contigs linked via paired-end reads, and 26 inserts were partially sequenced. In total, 2.5 Mb of metagenomic DNA sequence was assembled and manually edited from the 98 fosmids selected for sequencing.

Binning. MEGAN was used to determine the phylogenetic distribution of the first batch of 30,000 Sanger reads generated by the CSP program. BLASTX was used to compare all reads against the NCBI-NR ("non-redundant") protein database. Results of the BLASTX search were subsequently uploaded into MEGAN (6) for hierarchical tree constructions, which uses the BLAST bit-score to assign taxonomy, as opposed to using percentage identity.

Assembled metagenomic contigs were binned (classified) using PhyloPythia (7). Generic models for the ranks of domain, phylum, and class were combined with sample-specific models for the clades "uncultured γ-*Proteobacteriacea bacterium*" (WG-1), "uncultured *Lachnospiraceae bacterium*" (WG-2), and "uncultured *Erysipelotrichaceae bacterium*" (WG-3). The generic models represent all clades covered by two or more species at the corresponding ranks among the sequenced microbial isolates. The sample-specific models include classes for the dominant sample populations of WG-1, WG-2, and WG-3, as well as a class "Other." The sample-specific models for WG-1, WG-2, and WG-3 were each trained on sequence data obtained from contigs assembled from the metagenome, and fully sequenced fosmids identified using phylogenetic marker genes. Five sample-specific support vector machines were created by using fragments of lengths of 3, 5, 10, 15, and 50 kb. All input sequences were extended by their reverse complement before computation of the compositional feature vectors. The parameters *w* and *l* were both set to 5 for the sample-specific models. Thirty-three assembled contigs and one sequenced fosmid, assigned unambiguously through analysis of phylogenetic marker genes, were used for the training of WG-1 sample-specific model (a total of 388,452 bp). Sixteen assembled contigs and four sequenced fosmids assigned unambiguously through analysis of phylogenetic marker genes were used for the training of WG-2 sample-specific models (a total of 208,429 bp). Thirty assembled contigs assigned unambiguously through analysis of phylogenetic marker genes were used for the training of WG-3 sample-specific model (a total of 118,249 bp). Input fragments of a particular length were generated from the fosmids by using a sliding window with a step size of one-tenth of the generated fragment size (for

example, 5 kb for 50-kb fragments). For the class "Other," fragments from 340 sequenced isolates was used. The classifier consisting of the sample-specific and generic clade models were then applied to assign all fragments more than 1 kb of the sample. In case of conflicting assignments, preference was given to assignments of the sample-specific models. Results of this binning process were loaded into IMG/M-ER to allow independent analysis of the component populations. The resulting metabolic reconstruction for WG-2 is summarized in Fig. S4, with IMG gene object identifiers (oids) for tracking in IMG/M provided in Table S5.

1. Kang S, Denman SE, Morrison M, Yu Z, McSweeney CS (2009) An efficient RNA extraction method for estimating gut microbial diversity by polymerase chain reaction. *Curr Microbiol* 58:464–471.
2. Warnecke F, et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450:560–565.
3. Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: A program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* 20:2317–2319.
4. Cole JR, et al. (2003) Ribosomal Database Project (2003) The Ribosomal Database Project (RDP-II): Previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res* 31:442–443.
5. Chou H-H, Holmes MH (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* 17:1093–1104.
6. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
7. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 4:63–72.

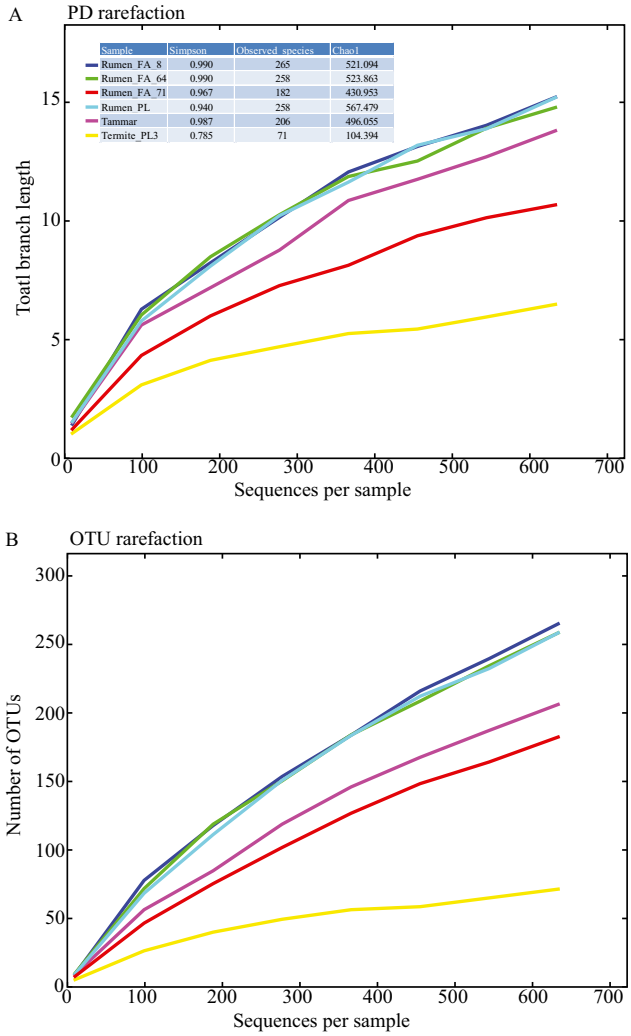

Fig. S1. Rarefaction analyses using phylogenetic diversity (PD) (*A*) and operational taxonomic unit (OTU) frequency (*B*) of *rrs* gene datasets obtained from the Tammar wallaby foregut, bovine rumen, and termite lumen microbiomes. In both panels, a 97% sequence identity threshold has been employed for the OTU constructions used in these analyses. (*A*) A "PD_whole tree" has been employed using QIIME and the PD value on the *y* axis represents the summation of the branch lengths from the phylogenetic trees constructed from the *rrs* gene sequences (Rumen_FA_8, dark blue; Rumen_PL, light blue; Rumen_FA_64, green; Tammar, purple; Rumen_FA_71, red and Termite_PL3, yellow). The α diversity measures (Simpson, observed species, and chao1 richness estimates) are also shown for each sample. (*B*) The same data as in *A* using the number of OTUs observed with each dataset as the *y* axis.

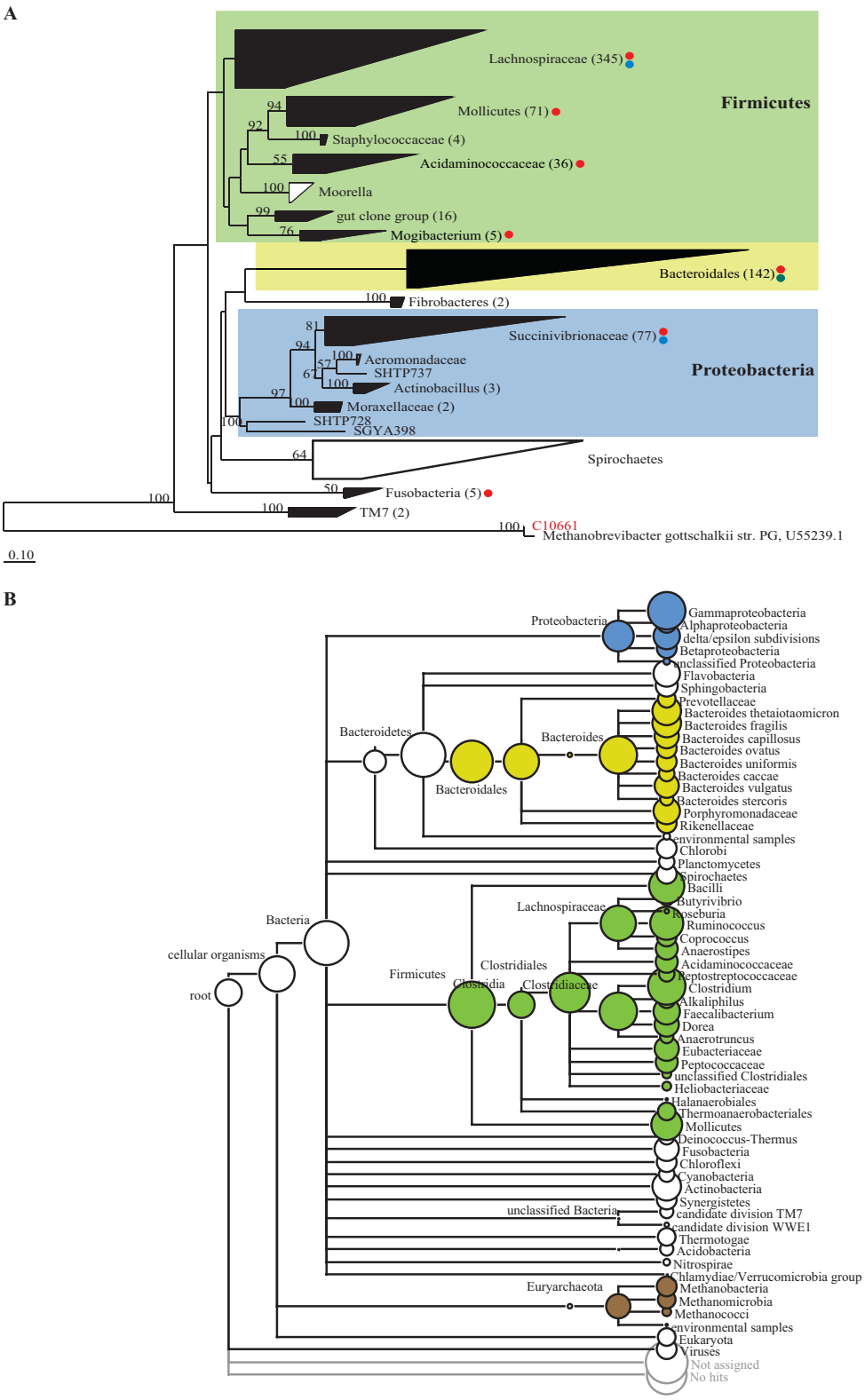

Fig. S2. Composition of the bacterial community analyzed. (*A*) Phylogenetic diversity of the Tammar wallaby foregut *rrs* sequences. From a PCR-based collection and the metagenomic libraries (small-insert and fosmid), 663 almost full-length and 51 partial to full 16S rRNA gene sequences representing six different phyla were analyzed using the maximum-likelihood algorithm. The total number of sequences contained within each grouping is noted in brackets. Red dots indicate where at least one group was represented in both the PCR collection and metagenomic libraries. Blue dots indicate where at least one group was represented by a sequenced fosmid clone. The phylogram was constructed from 1,289 unambiguously aligned nucleotide positions. (Scale bar, a sequence divergence of 10%.) Branching pattern confidence values greater than 50% are shown at nodes. See Figs. S1 and S3 and Table S2 for accession numbers and detailed phylogenetic analysis. (*B*) Phylogenetic diversity of metagenome sequences (~30 000) computed by MEGAN based on a BLASTX comparison. The size of the circles is scaled logarithmically to represent the number of reads assigned directly to the taxon.

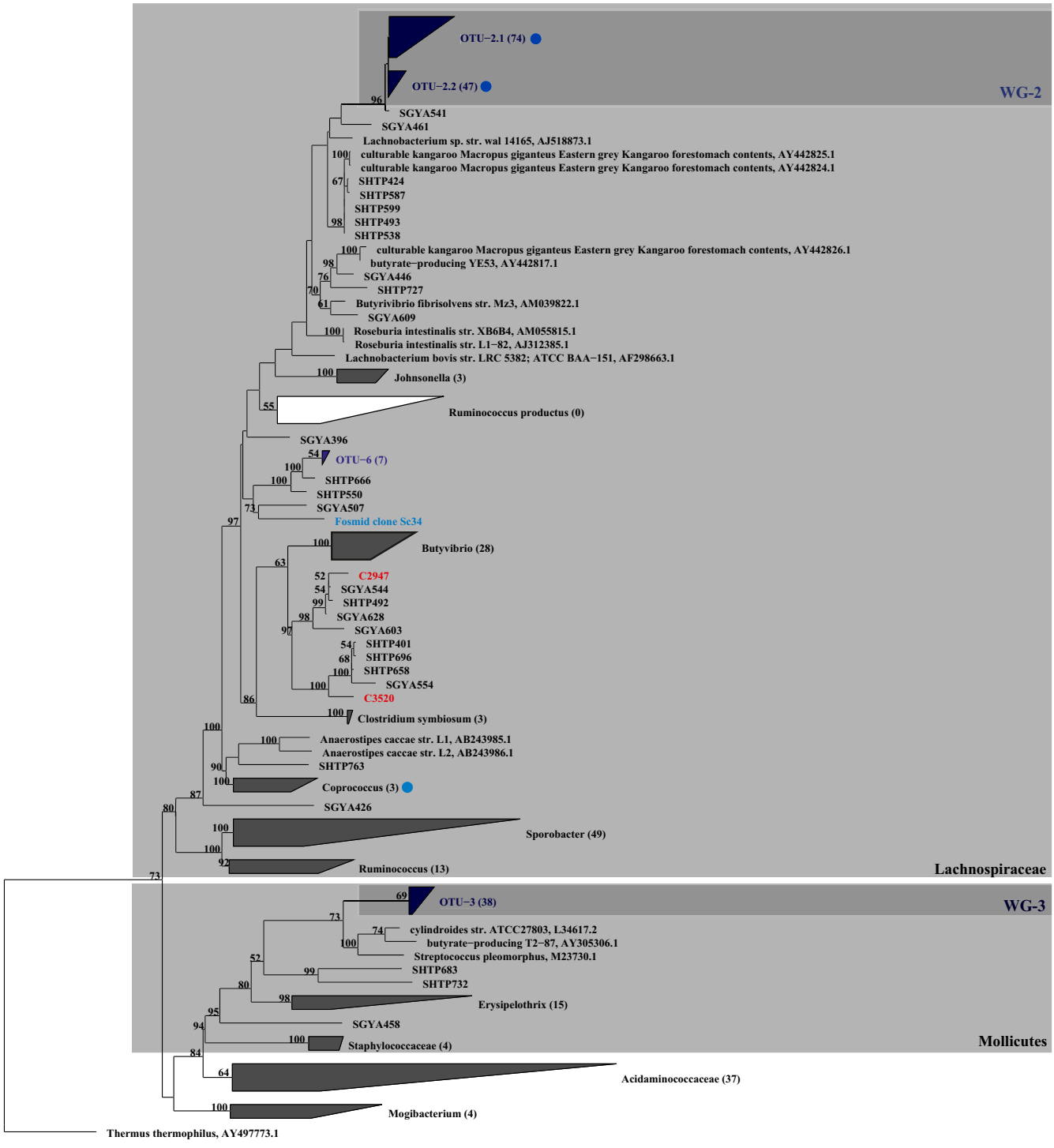

Fig. S3. Phylogenetic diversity of the Tammar wallaby foregut microbiota within the phylum Firmicutes. From a PCR-based collection and the metagenomic libraries (small-insert and fosmid), 663 almost full-length and 51 partial to full 16S rRNA gene sequences were analyzed using the maximum-likelihood algorithm (RAxML). The number of Tammar foregut community sequences within each grouping is given in brackets; dark blue shading denotes dinstict Tammar foregut community phylotypes; red text denotes sequences from the metagenome libraries. Light blue dots indicate where at least one OTU was represented by a sequenced fosmid clone. White shading denotes reference groups having no representation in this collection. The phylogram was constructed from 1,289 unambiguously aligned nucleotide positions. The scale bar represents a sequence divergence of 10%. Branching pattern confidence values greater than 50% are shown at nodes.

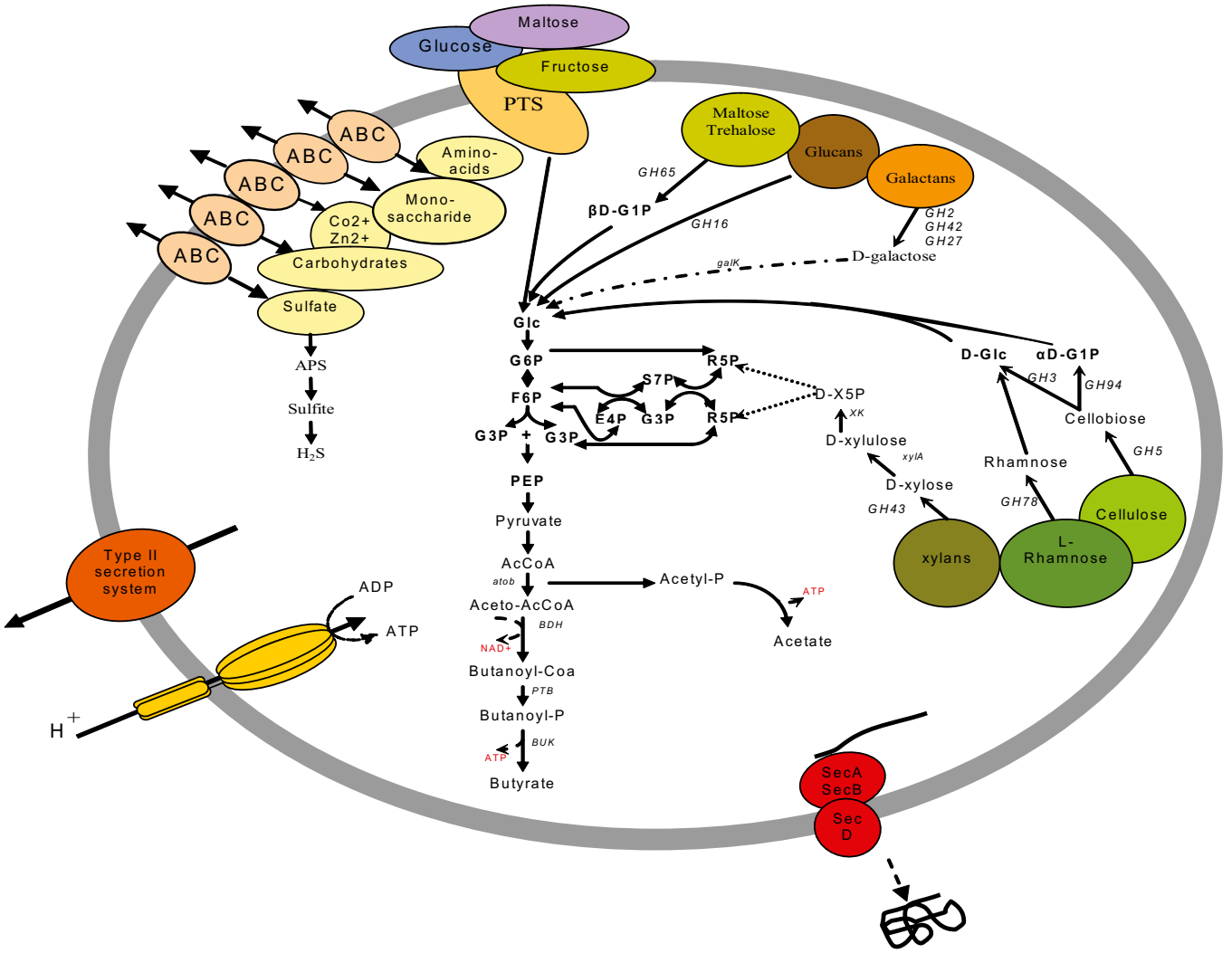

Fig. S4. Schematic representation of selected metabolic features of the WG-2 population as inferred from genomic comparisons (Table S5). Major metabolic pathways and energy-generating systems are shown. Similar shapes indicate similar functions. Broken lines indicate sections of pathways missing; partially broken lines indicate partial sections of pathway identified. Abbreviations: AcCoA, acetyl-CoA; acetyl-P, acetyl phosphate; AK, acetate kinase; APS, adenylylsulfate; atob, acetyl-CoA C-acetyltransferase; αD-G1P, α-D-Glucose 1-phosphate; BDH, butyryl-CoA dehydrogenase; BU.K., butyrate kinase; butanoyl-P, butanoyl phosphate; CL, citrate lyase; D-X5P, D-Xylulose 5-phosphate; E4P, erythrose-4-phosphate; F6P, fructose-6-phosphate; G3P, glyceraldehyde-3-phosphate; G6P, glucose-6-phosphate; galK, galactokinase; GH, glycoside hydrolase family (CAZy); PTB, phosphate butyryltransferase; PTS, Phosphotransferase system; R5P, pentose-phosphates; S7P, sedoheptulose-7-phosphate; XK, xylulokinase; xylA, xylose isomerize. Gene identification numbers (IMG gene object identifiers) can be found in Table S5.

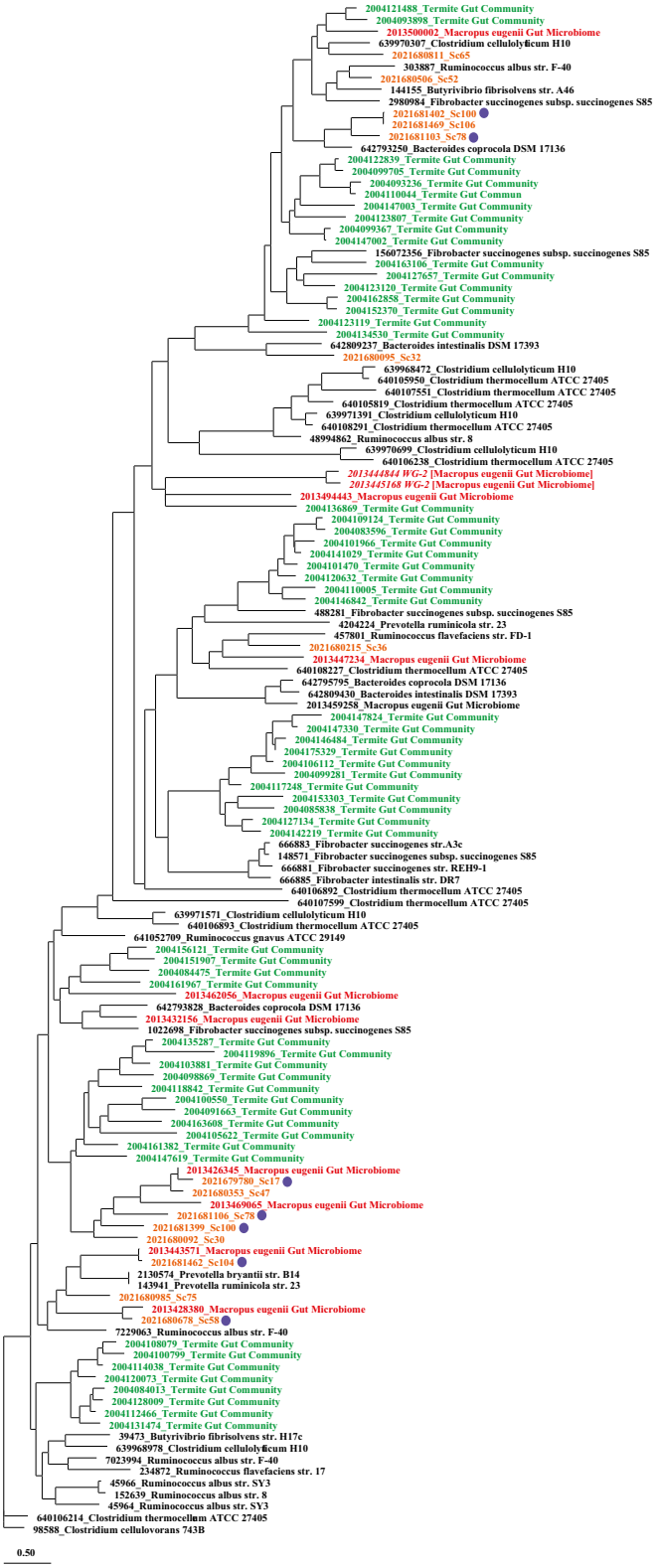

Fig. S5. Phylogenetic analysis of the glycoside hydrolase family 5 (GH5) diversity encoded by the Tammar wallaby foregut microbiome. GH5 sequences from the Tammar foregut metagenome are colored in red, GH5 sequences from the sequenced fosmid clones are colored in yellow, the termite gut metagenome is green, and various other sources in black. Purple dots denote GH5 sequence recovered from putative polysaccharide utilization loci gene clusters (Table S4). Metagenomic sequences and additional public sequences are identified by their Joint Genome Institute gene object identifier and GenBank GI number, respectively. The Pfam PF00150 (Cellulase – glycosyl hydrolase family 5) has a length of 378 residues, comprising domain sequences with an average length of 227 amino acids and 17% sequence identity.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (PDF)
Table S4 (PDF)
Table S5 (PDF)