ADAPTATION TO HERBIVORY BY THE TAMMAR WALLABY INCLUDES BACTERIAL AND GLYCOSIDE HYDROLASE PROFILES DIFFERENT TO OTHER HERBIVORES

P.B. Pope[1], S.E. Denman[1], M. Jones[1], S.G. Tringe[2], K. Barry[2], S.A. Malfatti[2], A.C. McHardy[3], J-F. Cheng[2], P. Hugenholtz[2], C.S. McSweeney[1], and M. Morrison[1,4*]

[1]CSIRO Livestock Industries, Queensland Bioscience Precinct, St Lucia, 4069 AUSTRALIA

[2]United States Department of Energy, Joint Genome Institute, Walnut Creek CA, 94598 USA

[3]Computational Genomics and Epidemiology Group, Max Planck Institute for Computer Science, Campus E1 4, 66123 Saarbruecken, GERMANY.

[4]The MAPLE Research Initiative, Department of Animal Sciences, The Ohio State University, Columbus OH 43210, USA

***Corresponding Author:** Mark Morrison
Queensland Bioscience Precinct
306 Carmody Road
St Lucia, QLD 4068
Australia
Phone: +61 4245 97323
Fax: +61 7 3875 7800
Email: Mark.Morrison@csiro.au

JULY 28, 2010

ACKNOWLEDGMENT

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No.
DE-AC02-05CH11231

DISCLAIMER

ADAPTATION TO HERBIVORY BY THE TAMMAR WALLABY INCLUDES BACTERIAL AND GLYCOSIDE HYDROLASE PROFILES DIFFERENT TO OTHER HERBIVORES

P.B. Pope[1], S.E. Denman[1], M. Jones[1], S.G. Tringe[2], K. Barry[2], S.A. Malfatti[2], A.C. McHardy[3], J-F. Cheng[2], P. Hugenholtz[2], C.S. McSweeney[1], and M. Morrison[1,4*]

[1]CSIRO Livestock Industries, Queensland Bioscience Precinct, St Lucia, 4069 AUSTRALIA

[2]United States Department of Energy, Joint Genome Institute, Walnut Creek CA, 94598 USA

[3]Computational Genomics and Epidemiology Group, Max Planck Institute for Computer Science, Campus E1 4, 66123 Saarbruecken, GERMANY.

[4]The MAPLE Research Initiative, Department of Animal Sciences, The Ohio State University, Columbus OH 43210, USA

***Corresponding Author:** Mark Morrison
Queensland Bioscience Precinct
306 Carmody Road
St Lucia, QLD 4068
Australia
Phone: +61 4245 97323
Fax: +61 7 3875 7800
Email: Mark.Morrison@csiro.au

Abstract

Metagenomic and bioinformatic approaches were used to characterize plant biomass conversion within the foregut microbiome of Australia's "model" marsupial, the Tammar wallaby (*Macropus eugenii*). Like the termite hindgut and bovine rumen, key enzymes and modular structures characteristic of the "free enzyme" and "cellulosome" paradigms of cellulose solubilisation remain either poorly represented or elusive to capture by shotgun sequencing methods. Instead, multi-gene polysaccharide utilization loci (PULs)-like systems coupled with genes encoding β-1,4-endoglucanases and β-1,4-endoxylanases - which have not been previously encountered in metagenomic datasets were identified; as well as a diverse set of glycoside hydrolases targeting non-cellulosic polysaccharides. Furthermore, both *rrs* gene and other phylogenetic analyses confirmed that "novel" clades of the Lachnospiraceae, Bacteroidales and Gammaproteobacteria are predominant in the Tammar foregut microbiome. Nucleotide composition-based sequence binning facilitated the assemblage of more than 2 Megabase pairs of genomic sequence for one of the novel Lachnospiraceae clades (WG-2). These analyses show that WG-2 possesses numerous glycoside hydrolases targeting non-cellulosic polysaccharides. These collective data demonstrate that Australian macropods not only harbor "unique" bacterial lineages underpinning plant biomass conversion, but their repertoire of glycoside hydrolases is distinct from those of the microbiomes of higher termites and the bovine rumen.

\body

INTRODUCTION

Australia possesses the largest share of the world's extant marsupial species, which diverged from other eutherian mammals approximately 150 million years ago. Probably the most widely recognized members of this group are the macropods (kangaroos and wallabies). The macropods also evolved in geographical isolation of other eutherian herbivores and although they are often compared to ruminants, the various macropod species show a wide range of unique adaptations to herbivory. These differences include their dentition and mastication of food, as well as the anatomical adaptations of the forestomach that supports a cooperative host-microbe association that efficiently derives nutrients from plant biomass rich in lignocellulose (1). Compared to ruminant species, the hydrolytic and fermentative processes these microbes provide must be relatively rapid because of the continuous transit of plant biomass through the herbivore gut (2, 3). There is also a widespread belief - developed from several studies during the late 1970's - that Australian macropods generate less methane during feed digestion than ruminant herbivores (4, 5); indicative of some novel host and/or microbe adaptations of the macropods to herbivory. Indeed, the limited studies published to date suggest the foregut microbiomes of macropods possess "novel" protozoal, bacterial and archaeal microorganisms (6-8), however very little is currently known about the genetic potential and structure-function relationships intrinsic to these microbiomes.

Metagenomics offers new opportunities to interrogate and understand this interesting host-microbe association. We present here a compositional and comparative analysis of metagenomic data pertaining to plant biomass hydrolysis by the foregut microbiome of Australia's model marsupial: the Tammar wallaby *(Macropus eugenii)*. Several novel bacterial lineages were identified and nucleotide composition-based sequence binning using PhyloPythia facilitated the production of a 2.3 Mbp assemblage of DNA representing one of the novel Lachnospiraceae clades present in this community. Further *in silico* analysis revealed this clade harbours numerous putative glycoside hydrolases specifically targeting the side chains attached to non-cellulosic polysaccharides.

RESULTS AND DISCUSSION

Microbial diversity resident in the Tammar wallaby foregut. An inventory of the various metagenomic resources created and analysed as part of this study are summarized in **Table S1**. The *rrs* gene library is comprised of 663 bacterial sequences and included 293 phylotypes (using a 99% sequence identity threshold) and rarefaction analysis showed the bovine and macropod datasets afforded a similar degree of coverage of the biodiversity present in these microbiomes. The overall community profile at a phylum-level is similar to that of other vertebrate herbivores with representatives of the Firmicutes and Bacteroidetes predominant (**Fig. S2a**, **Table S2**). However, the majority of these phylotypes were only distantly related to any of the cultivated species from other gut microbiomes (**Table S2**). Furthermore, the comparison of these datasets via unweighted measures of β diversity, UniFrac analysis and OTU network maps clearly showed host-specificity, with only a small number of OTU's shared between the bovine and macropod microbiomes, and no OTU's shared with the termite sample (Fig. 1). We were also able to separate the macropod *rrs* gene library data with respect to time of collection, which revealed that the microbiome appeared to be more diverse in spring, most likely due to the availability of forb species during spring offering a greater amount of soluble carbohydrates, as compared to highly lignocellulosic biomass present in drier times of the year (Fig. 1; **Table S2)**. There were also five distinctive phylotypes identified from the *rrs* gene libraries: two of these were assigned as deeply branching, novel members of the gamma-subdivision of Proteobacteria (hereafter referred to as Wallaby Group-1, WG-1); two more were positioned as a deeply branching and novel lineage within the Lachnospiraceae (hereafter referred to as Wallaby Group 2, WG-2) and the last of these was a novel member of the Erysipelotrichaceae (Mollicutes), and is hereafter referred to as Wallaby Group-3, (WG-3, **see Table S2 and Fig. S3**). The archaeal populations have already been described in an earlier study (8) and are considerably smaller than those typically encountered in ruminants, and might be a key reason explaining the apparent differences between ruminants and macropods in terms of methane production during feed digestion (2, 4, 5).

Similar conclusions were drawn from phylogenetic analysis of the Sanger shotgun sequence data. First, MEGAN (9) was used to perform a phylogenetic assignment of the first round of metagenomic data generated, which represented approximately 30% of the total data produced (**Fig. S2b**). The majority of these reads were assigned to the

Firmicutes, Bacteroidetes and the gamma-subdivision of the Proteobacteria. We subsequently developed a collaborative partnership with the McHardy group and used the composition-based classifier PhyloPythia (10) to examine the complete dataset, once it was produced. The fosmid libraries produced as part of this study were used to provide ~2.5 Mbp of training sequence for Phylopythia, which resulted in the classification of 76% of the contigs to at least the phylum level (Table 1). Again, the assignments favored the Firmicutes, Bacteroidetes and gamma-subdivision of the Proteobacteria; and confirmed the predominance of the WG-1, WG-2 and WG-3 populations in the metagenomic data (Table 1). Indeed these three groups, which comprise ~34% of the sequences that comprise the *rrs* gene libraries (**Table S2**), also accounted for ~22% of the total PhyloPythia assignments. A small number of reads were also assigned to the Euryarchaeota and in particular, *Methanobrevibacter* sp.; consistent with the small population size of archaea measured for these same animals by (8).

Despite these encouraging results, approximately 60% of the Sanger reads subjected to MEGAN analysis and 61% of the Phylopythia assignments could not be extended deeper than an *Order*-level of classification, with ~20% having no assignments at any level. This "shallow" level of binning by both methods confirms the wallaby foregut microbiome is comprised of "novel" bacterial lineages, with only limited similarity to the (meta)genomic data derived from other microbial habitats and cultured isolates.

The Tammar foregut microbiome possesses a different repertoire of GH genes and related modules, compared to other herbivore microbiomes. The Tammar wallaby is a small macropod (4-10 kg) and primarily utilizes grasses and forbs as its principal source of energy nutrition (2). For many months of the year, such plant material is characteristically rich in lignocellulose and non-cellulosic polysaccharides. The metagenomic data was subjected to automated annotation using JGI-DOE's integrated microbial genomes with microbiome samples (IMG/M) system; then select functional categories were manually compared to the global HMMs available via pfam. These analyses recovered 600 genes and/or modules from 53 different CAZy families (11) (**Table S4**); but relatively few of these produced strong matches with endo- or exo-acting β-1,4-glucanases. Only 24 GH5 β-1,4-endoglucanases were identified from the metagenomic data, along with a smaller number of gene modules

assigned to the GH6, GH8, GH9 and GH74 families (**Table S4**). In addition to these presumptive "cellulases", the metagenomic data produced 25 sequences matching GH94 (cellobiose phosphorylase) catalytic modules. The number of "xylanase" genes identified in the metagenomic dataset were evenly distributed among the GH10 (n=14), GH26 (n=13) and GH43 (n=19) families (**Table S4**). An additional three GH10 genes were retrieved from the fosmid library by functional screens and binned to the Bacteroidetes (**Table S5**). Interestingly the GH11 xylanases, which are found in abundance among members of the Firmicutes, especially *Clostridium* and *Ruminococcus* spp., as well as specialist cellulolytic bacteria from other gut microbiomes, were absent from our datasets.

Comparative analysis of the repertoire of GH families recovered from the Tammar, termite hindgut and bovine rumen metagenomes revealed some interesting similarities and differences. The GH5 "cellulases" were numerically most abundant in the wallaby and termite metagenomes, with less representation of the GH9 family (Table 2). In contrast, the bovine metagenomic dataset was more evenly balanced with respect to these two GH families ((12, 13); Table 2). Similar to the rumen, the wallaby foregut microbiome possessed a large number of reads matching GH families specific for xylooligosaccharides and the side chains attached to non-cellulosic polysaccharides (Table 2). The most abundant were GH1, GH2 and GH3 β-glycosidases, as well as matches with GH51 and GH67 enzymes which typically target glucuronic acid and arabinose-containing side chains, respectively. The Tammar metagenome also contained a range of carbohydrate active enzymes targeting pectic polysaccharides, plant pigments, gums, glycolipids and other glycosides; including GH78 rhamnosidases, CE8 pectin methylesterases, several GH28 rhamnogalacturonases, and pectate lyases (Table S4). These findings were not entirely unexpected given the dietary profiles of the macropod (predominantly grass and forbs, with a small amount of a commercial pellet mix) compared to termites (wood); and also partially explains the higher abundance of GH genes that catalyse the hydrolysis of the side chains of non-cellulosic plant polysaccharides in the grass/legume feeding herbivores when compared to wood-eating termites (13).

However, and despite the differences in nutritional ecology, gut anatomy and microbiome structure, probably the most notable observation drawn from all these

datasets is the virtual absence of genes encoding GH6, GH7 and GH48 β-1,4-exoglucanases (Table 2), which are essential in virtually all cultured bacteria and fungi for cellulose solubilisation; as well as the dearth of cellulosome-associated modules, such as cohesins and dockerins (Table S4). Although the wallaby metagenome dataset does contain 42 Type I dockerin modules, all these modules were linked to hypothetical sequences of unknown function; with no examples linked to recognized GH catalytic modules, other carbohydrate-active enzymes, or serpins. Such findings suggest there is still much to learn about cellulose hydrolysis, and dockerin-cohesin-mediated complex assemblies, in gut microbiomes.

Identification of novel PUL-like gene clusters associated with "cellulase" genes in the sequenced fosmid clones. There were 33 fosmids selected for 454 pyrosequencing on the basis that their inserts encode gene product(s) resulting in carboxymethylcellulase (CMCase) or xylanase activity visualized in plate screen assays. PhyloPythia assigned the majority of the scaffolds produced from these clones to the Bacteroidales or Lachnospiraceae (**Table S5**). Twelve of these scaffolds possess genes encoding a GH5 catalytic module, two more encode a gene with a GH9 catalytic module and one encodes a gene with a GH6 catalytic module. Interestingly, half of the scaffolds assigned to the Bacteroidales also possessed genes homologous to the polysaccharide utilization loci (PULs) present in the genomes of *Bacteroides* and related genera (14-18). The presumptive PUL-like gene arrangement borne by one of these fosmids (annotated in **Table S5** as part of scaffold 78) is shown as an example in Fig. 2, along with a hypothetical functional model of the cluster. In brief detail, the Pul-like gene cluster consists of an AraC-like regulatory protein, a putative acetylxylan esterase and two genes with homology to the *B. thetaiotaomicron susC (tonB)* and *susD* genes. These latter two genes were initially defined as part of the starch utilization system (*sus*) of *B. thetaiotaomicron* (19, 20). The SusC protein is a *tonB*-dependent receptor family member, a group of outer membrane-spanning proteins that can import solutes and macromolecules into the periplasm (21, 22); the SusD protein coordinates polysaccharide binding at the cell surface (20). Two genes located directly downstream from the *susC* and *susD* homologs were predicted to be outer membrane lipoproteins and therefore might play a role similar to the *B. thetaiotaomicron* SusE and SusF proteins whose functional role is currently unknown. The remaining six genes in this cluster encode putative glycoside

hydrolases and a putative inner-membrane bound "sugar transporter". Although PULs were not readily assembled from our Sanger sequence data; there were 34 *susC* and 38 *susD* genes identified in the dataset. For these reasons, we propose that the *sus*-like PULs represent a key adaptation to growth on cellulose and other polysaccharides by the large number of Bacteroidetes resident in the wallaby foregut. Interestingly, *sus* gene homologs were not identified in the termite hindgut and bovine rumen data; presumably due to the lower representation of Bacteroidetes in the termite and the short read lengths in the bovine dataset.

Phylopythia-supported metabolic reconstruction of the WG-2 population. Phylopythia supported a 2.3 Mbp assemblage of metagenome fragments assigned to the WG-2 population (**Fig. S5; Table S3**). The current assemblage includes 20 different families of carbohydrate-active enzymes, principally involved with the hydrolysis of non-cellulosic polysaccharides and pectin. However, none of the sequences encoding dockerin modules were assigned to WG-2, suggesting no cellulosome complex assembly by this population. The assemblage includes genes encoding homologs of GH1, GH2, GH3, GH27 and GH42 catalytic modules, as well as several GH5 endoglucanases and GH94 cellobiose phosphorylases. Five GH43 arabinoxylosidases were also assigned to WG-2, and several acetyl esterases genes contiguous with GH78 rhamnosidases were also assigned to WG-2 (**Table S3**). Interestingly, arabinose-rich rhamnogalacturonan side chains have been speculated to play an essential role for some plant species to tolerate severe desiccation (23, 24). Given many of Australia's native plant species are drought tolerant or drought resistant, WG-2 might have evolved to specialise in the hydrolysis and use of these types of poly- and oligosaccharides for growth. Indeed, Phylopythia also assigned genes encoding xylose isomerase and xylulokinase enzymes to the WG-2 assemblage, as well as acetate and butyrate kinases. From these data we propose that the WG-2 population plays a quantitatively important role in both the degradation and fermentation of the pentoses derived from non-cellulosic polysaccharides, and produces acetate and butyrate as fermentation end products.

Australia's flora and fauna are recognized throughout the world for their unique attributes in diversity, form and function, but our understanding of their evolutionary adaptations for niche occupation has been compromised because we had virtually no

understanding of their gut microbiomes, which contribute greatly to the nutrition and well-being of these animals. Our metagenomic analyses of the Tammar wallaby foregut microbiome clearly shows these animals are the host for "novel" bacterial lineages that are numerically predominant within the microbiome. For instance, the WG-2 lineage appears to play a key role in the deconstruction of non-cellulosic poly- and oligosaccharides by producing a large number of enzymes targeting both heteroxylans and pectins. Furthermore, the functional screening of the fosmid libraries for "cellulases" and "xylanases" recovered clones assigned to the Bacteroidetes encoding PUL-like gene clusters; including *susC* and *susD* gene homologs linked with GH5 and/or GH10 genes. Such findings distinguish the Tammar wallaby foregut microbiome from that of the bovine rumen (predominantly Clostridiales and Prevotellas) and the termite hindgut (Fibrobacteres and Spirochetes). The collective findings from this and other metagenomic studies also still need to be reconciled with the extensive literature developed from the biochemical, molecular and genomic analyses of specialist gut bacteria and fungi; which have created the cellulosome and free enzyme paradigms of cellulose solubilisation. These paradigms are underpinned by a restricted number of known GH families, which remain poorly represented in metagenomic data. Much still remains to be learned about the structure-function relationships of these interesting microbiomes.

MATERIALS AND METHODS

Wallaby sampling. The eight adult females (aged between 1.5 and 4 years) sampled for this study were all from the same colony maintained near Canberra, Australia. Three animals were sampled in November 2006 (late spring) and another 5 in May 2007 (late autumn). During this period the animals were provided free range access to pastures composed predominantly of Timothy Canary grass (*Phalaris angusta*) and were also provided with a commercial pellet mix containing wheat, bran, pollard, canola, soy, salt, sodium bicarbonate, bentonite, lime and a vitamin premix (Young Stockfeeds, NSW, Australia). Animals were euthanized with an overdose of pentobarbitone sodium (CSIRO Sustainable Ecosystems Animal Ethics Approval Number 06-20) and foregut contents were either transferred to sterile containers and immediately frozen at -20°C, or mixed 1:1 with phenol:ethanol (5%:95%).

Cell dissociation and DNA extraction. Prior to 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 5-10 g of the pooled samples was centrifuged at 14 000 rpm for 2 minutes, and the pellet was resuspended in dissociation buffer and subjected to a dissociation procedure described by (25) (**details provided in *SI text***).

The cell pellets (~200 mg wet-weight) were resuspended in 700 µl TE buffer and incubated at 75^{o}C for 10 minutes to inactivate nucleases. Cell lysis was performed by adding lysozyme (1mg/ml)/ mutanolysin (20 U) and achromopeptidase (1 mg/ml) to these cell suspensions and incubation at 37^{o}C for 90 minutes. Then SDS was added to give a final concentration of 1 % (w/v) and 0.20 mg proteinase K was also added, and the mixture was incubated at 55^{o}C for 90 minutes. Next, NaCl and CTAB were added to give final concentrations of 0.7 M and 2% (w/v) respectively, and the mixture was incubated at 70^{o}C for 10 minutes. Following phenol:chloroform:isoamylalcohol and chloroform extractions, the DNA was precipitated with 2 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 metagenomic DNA samples extracted from T1 and T2, 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') (12) (**details provided in *SI text***). Similar attempts produced archaeal *rrs* gene libraries with results described in (8). A total of 663 near-complete bacterial *rrs* gene sequences passed the quality and chimera filters and were used in the subsequent analyses (**details provided in *SI text***).

Phylogenetic analysis of 16S rRNA gene sequences. The 663 sequences were aligned using the NAST aligner (26) and imported into an ARB database with the same alignment (http://greengenes.lbl.gov/) (27). Fifty-one partial and near complete 16S sequences were extracted from the Tammar metagenomic data set aligned using

NAST aligner and also imported into ARB (28). Sequences were initially assigned to phylogenetic groups using the ARB Parsimony insertion tool. Phylogenetic trees (**Fig. S2 and Fig. S3**) were constructed from masked ARB alignments (to remove ambiguously alignable positions) using RAxML (29) and bootstrap analysis using parsimony and neighbour-joining was performed using 100 replicates. The phylum-level trees (**Fig. S3**) were reconstructed using TREE-PUZZLE (30) in ARB. The *rrs* gene sequences from the two libraries were assigned to clusters (operational taxonomical units; OTUs) at 97% and 99% sequence identity thresholds using the DOTUR package (31) and comparisons at an OTU definition was calculated as a percentage using SONS (32) (Table S2). Additional phylogenetic comparisons and diversity estimates were performed using the QIIME package (Quantitative Insights Into Microbial Ecology) (33), with OTUs at the 97% sequence identity threshold used. Sample heterogeneity was removed by rarefaction prior to comparison of Tammar *rrs* gene sequences with rumen and termite samples. The OTU network maps were generated using QIIME and visualised with Cytoscape (34). In addition, alpha diversity (PD_Whole_Tree, observed species count and Chao1 richness estimators) and beta diversity (unifrac weighted and unweighted) metrics along with rarefaction plots were also calculated using QIIME.

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 2-4 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. Subsequent sequences were assembled with the Paracel Genome Assembler (PGA version 2.62, www.paracel.com) (**details provided in *SI text***).

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 2 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.

Gene prediction. Putative genes in the Tammar Wallaby gut microbiome metagenome were called with GeneMark (35) and putative genes in the fosmid assemblies were called with a combination of MetaGene (36) and BLASTx. All called genes were annotated via the IMG/M-ER annotation pipeline and loaded as independent data sets into IMG/M-ER (37) (http://img.jgi.doe.gov/cgi-bin/m/main.cgi), a data-management and analysis platform for genomic and metagenomic data based on IMG (38).

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 (9) 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 (10). Generic models for the ranks of domain, phylum and class were combined with sample-specific models for the clades "uncultured gamma-Proteobacteria bacterium" (WG-1), "uncultured Lachnospiraceae bacterium" (WG-2) and "uncultured Erysipelotrichaceae bacterium" (WG-3) (**details provided in *SI text***).

Glycoside hydrolases (GH) and carbohydrate-binding modules (CBM): annotation and phylogenetic analysis. Searches for GHs and CBMs were performed as described by (12). Briefly, database searches were performed using HMMER hmmsearch with pfam_Is HMMs (full length models) to identify complete matches to the family, which were named in accordance with the CAZy nomenclature scheme (11). All hits with E-values less than 10^{-4} were counted and their sequences further

analysed. For those GH and CBM families for which there is currently no Pfam HMM, the representative sequences selected from the CAZy website and described by (12) were used in BLAST searches of the metagenomic data to identify these GH and CBM families. An E-value cutoff of 10^{-6} was used in these searches. For phylogenetic analysis of selected GH families, sequence alignments were first produced using HMMER hmmalign and to the corresponding pfam HMM; then a protein maximum likelihood program (PROML) used with the Jones-Taylor-Thornton probability model of change between amino acids was applied to these data.

Identification of fosmid clones bearing GH gene(s). Fosmid clones bearing β-1,4-endoglucanase and/or β-1,4-xylanase activity were detected by plating the *E. coli* library on LB-chloramphenicol agar plate medium containing either 0.2% (w/v) carboxymethylcellulose or birchwood xylan (Sigma). Approximately 20,000 recombinant strains were plated in a 384-well format and incubated overnight at 37°C. The plates were then stained with Congo red dye and de-stained with 1M NaCl to reveal zones of hydrolysis. Positive colonies were isolated and reexamined to confirm activity. Twenty-seven fosmid clones positive for carboxymethylcellulose hydrolysis and six positive for xylan hydrolysis were selected for 454 pyrosequencing and assembly.

ACKNOWLEDGEMENTS

The Tammar wallaby project is partially supported by CSIRO's Office of the Chief Executive (OCE) Science Leader program (awarded to MM), a CSIRO OCE Postdoctoral Fellowship (awarded to PBP), and the United States Department of Energy-Joint Genome Institute Community Sequencing Program. We are especially grateful to the support from Lyn Hind (CSIRO Australia) who assisted in sample collection. This work was performed in part under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

AUTHOR INFORMATION

The whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession number ADGC00000000. 16S rRNA gene sequences are deposited under the accession numbers GQ358225-GQ358517.

REFERENCES

1. Hume, I. D. (1984) Microbial fermentation in herbivorous marsupials. *BioScience* 34: 435-440.

2. Smith, J. A. (2009) Macropod nutrition. *Vet. Clin. Exot. Anim.* 12: 197–208

3. Flint, H. J. (1997) The rumen microbial ecosystem - some recent developments. *Trends Microbiol.* 5: 483-488.

4. Kempton, T. J., Murray, R. M., & Leng, R. A. (1976) Rates of production of methane in the grey kangaroo and sheep. *Aust. J. Biol. Sci.* 29: 209-214.

5. Engelhardt, W., Wolter, S., & Lawrenz, H. (1978) Production of methane in two non-ruminant herbivores. *Comp. Biochem. Physiol.* 60: 309-311.

6. Dehority, B. A. (1996) A new family of entodiniomorph protozoa from the marsupial forestomach, with descriptions of a new genus and five new species. *J. Eukaryot. Microbiol.* 43: 285-297.

7. Ouwerkerk, D., *et al.* (2005) Characterization of culturable anaerobic bacteria from the forestomach of an eastern grey kangaroo, *Macropus giganteus*. *Lett. Appl. Microbiol.* 41: 327-333.

8. Evans, P. N., *et al.* (2009) Community composition and density of methanogens in the foregut of the tammar wallaby (*Macropus eugenii*). *Appl. Environ. Microbiol.* 75: 2598-2602.

9. Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007) MEGAN analysis of metagenomic data. *Genome Res* 17: 377-386.

10. McHardy, A. C., *et al.* (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Meth.* 4: 63-72.

11. Cantarel, B. L., *et al.* (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* 37: 233-238.

12. Warnecke, F., *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450: 560-565.

13. Brulc, J. M., *et al.* (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc. Natl. Acad. Sci. USA* 106: 1948-1953.

14. Bjursell, M. K., Martens, E. C., & Gordon, J. I. (2006) Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. *J. Biol. Chem.* 281: 36269-36279.

15. Xu, J., *et al.* (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* 299: 2074-2076.

16. Xu, J., *et al.* (2007) Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol.* 5: e156.

17. Xie, G., *et al.* (2007) Genome Sequence of the Cellulolytic Gliding Bacterium Cytophaga hutchinsonii. *Appl. Environ. Microbiol.* 73: 3536-3546.

18. Martens, E. C., Koropatkin, N. M., Smith, T. J., & Gordon, J. I. (2009) Complex glycan catabolism by the human gut microbiota: The bacteroidetes Sus-like paradigm. *J Biol. Chem.* 284: 24673-24677.

19. Cho, K. H. & Salyers, A. A. (2001) Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by Bacteroides thetaiotaomicron. *J. Bacteriol.* 183: 7224-7230.

20. Koropatkin, N. M., Martens, E. C., Gordon, J. I., & Smith, T. J. (2008) Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* 16: 1105-1115.

21. Ferguson, A. D. & Deisenhofer, J. (2002) TonB-dependent receptors-structural perspectives. *Biochim. Biophys. Acta.* 1565: 318-332.

22. Schauer, K., Rodionov, D. A., & de Reuse, H. (2008) New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem. Sci.* 33: 330-338.

23. Moore, J. P., *et al.* (2006) Response of the leaf cell wall to desiccation in the resurrection plant *Myrothamnus flabellifolia*. *Plant Physiol.* 141: 651-662.

24. Moore, J. P., Farrant, J. M., & Driouich, A. (2008) A role for pectin-associated arabinans in maintaining the flexibility of the plant cell wall during water deficit stress. *Plant Signal Behav.* 3: 102-104.

25. Kang, S., *et al.* (2009) An efficient RNA extraction method for estimating gut microbial diversity by polymerase chain reaction. *Curr. Microbiol.* 58: 464-471.

26. DeSantis, T. Z., Jr., *et al.* (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* 34: 394-399.

27. DeSantis, T. Z., *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72: 5069-5072.

28. Ludwig, W., *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.* 32: 1363-1371.

29. Stamatakis, A., Ludwig, T., & Meier, H. (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456-463.

30. Schmidt, H. A., Strimmer, K., Vingron, M., & von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18: 502-504.

31. Schloss, P. D. & Handelsman, J. (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Appl. Environ. Microbiol.* 71: 1501-1506.

32. Schloss, P. D. & Handelsman, J. (2006) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl. Environ. Microbiol.* 72: 6773-6779.

33. Caporaso, J. G.*, et al.* (2010) QIIME allows integration and analysis of high-throughput community sequencing data. *(Submitted)*.

34. Shannon, P.*, et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13: 2498-2504.

35. Borodovsky, M., Mills, R., Besemer, J., & Lomsadze, A. (2003) Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr. Protoc. Bioinformatics* Chapter 4: Unit 4.5.

36. Noguchi, H., Park, J., & Takagi, T. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34: 5623-5630.

37. Markowitz, V. M.*, et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 36: 534-538.

38. Markowitz, V. M.*, et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* 34: 344-348.

39. Allgaier, M., *et al*. (2010) Targeted Discovery of Glycoside Hydrolases from a Switchgrass-Adapted Compost Community. *PloS ONE.* 5: e8812.

FIGURE LEGENDS

Figure 1. OTU network map showing OTU interactions between all rarefied samples from the Tammar wallaby (spring and autumn), rumen and termite. Lines radiating from samples Rumen _FA_8, Rumen_FA_64, Rumen_FA_71 and

Rumen_PL are coloured blue (fibre associated fraction and pooled liquid associated respectively from study (13)), Termite_PL3 coloured red (termite lumen study (12)) and Tammar_Spring and Tammar_Autumn colured green (this study) are weighted with respect to contribution to the OTU. OTU size is weighted with respect to sequence counts within the OTU. Insert shows the first two principal coordinate axes (PCoA) for the unweighted UniFrac analysis coloured by host animal; Rumen (FA_8, ■; FA_64, ●; FA_71, ♦; PL, ◄) blue; Termite (▲), red and Tammar (Spring, ►; Autumn, ▼) green. For complete inventory and comparisons between the two Tammar wallaby sample dates at an OTU definition (SONS analysis) see **Table S2.**

Figure 2. Gene arrangement in the Bacteroidales-affiliated fosmid and a hypothetical model of polysaccharide-adhesion and hydrolysis coordinated by this gene cluster. a. PhyloPythia affiliated the fosmid clone from which scaffold 78 is derived to the order Bacteroidales, as described in the text. The putative PUL gene cluster consists of an AraC family transcriptional regulator (geneA), an acetylxylan esterase (geneB), *susC* and *susD* gene homologs (genes C and D, respectively) and two genes encoding outer membrane-targeted lipoproteins (genes E and F). Genes G, H and I encode proteins containing GH5, GH26 and GH43 catalytic modules, respectively. Gene J encodes a putative inner-membrane bound "sugar transporter" followed by genes K and L, which encode proteins containing GH5 and GH94 catalytic modules, respectively. **b.** The hypothetical model predicts that polysaccharides are bound by the outer membrane-associated components, principally via the SusD homolog in a complex with the SusC, and the two lipoproteins. The GH5-containing proteins generate oligosaccharides, which are transported across the outer membrane, principally via the protein complex described above. These oligosaccharides may be further hydrolyzed by periplasmic GHs encoded by genes G, and I. The glycoside sugar transporter encoded by gene J, transports the hydrolysis products to the cytoplasm before terminal phosphorolytic cleavage by the GH94 glycoside phosphorylase (encoded by gene L).