LETTERS

Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite

Falk Warnecke[1]*, Peter Luginbühl[2]*, Natalia Ivanova[1], Majid Ghassemian[2], Toby H. Richardson[2]†, Justin T. Stege[2], Michelle Cayouette[2], Alice C. McHardy[3]†, Gordana Djordjevic[2], Nahla Aboushadi[2], Rotem Sorek[1], Susannah G. Tringe[1], Mircea Podar[4], Hector Garcia Martin[1], Victor Kunin[1], Daniel Dalevi[1], Julita Madejska[1], Edward Kirton[1], Darren Platt[1], Ernest Szeto[1], Asaf Salamov[1], Kerrie Barry[1], Natalia Mikhailova[1], Nikos C. Kyrpides[1], Eric G. Matson[5], Elizabeth A. Ottesen[6], Xinning Zhang[5], Myriam Hernández[7], Catalina Murillo[7], Luis G. Acosta[7], Isidore Rigoutsos[3], Giselle Tamayo[7], Brian D. Green[2], Cathy Chang[2], Edward M. Rubin[1], Eric J. Mathur[2]†, Dan E. Robertson[2], Philip Hugenholtz[1] & Jared R. Leadbetter[5]*

From the standpoints of both basic research and biotechnology, there is considerable interest in reaching a clearer understanding of the diversity of biological mechanisms employed during lignocellulose degradation. Globally, termites are an extremely successful group of wood-degrading organisms[1] and are therefore important both for their roles in carbon turnover in the environment and as potential sources of biochemical catalysts for efforts aimed at converting wood into biofuels. Only recently have data supported any direct role for the symbiotic bacteria in the gut of the termite in cellulose and xylan hydrolysis[2]. Here we use a metagenomic analysis of the bacterial community resident in the hindgut paunch of a wood-feeding 'higher' *Nasutitermes* species (which do not contain cellulose-fermenting protozoa) to show the presence of a large, diverse set of bacterial genes for cellulose and xylan hydrolysis. Many of these genes were expressed *in vivo* or had cellulase activity *in vitro*, and further analyses implicate spirochete and fibrobacter species in gut lignocellulose degradation. New insights into other important symbiotic functions including H_2 metabolism, CO_2-reductive acetogenesis and N_2 fixation are also provided by this first system-wide gene analysis of a microbial community specialized towards plant lignocellulose degradation. Our results underscore how complex even a 1-μl environment can be.

All known termite species form obligate, nutritional mutualisms with diverse gut microbial species found nowhere else in nature[3]. Despite nearly a century of study, however, science still has only a meagre understanding of the exact roles of the host and symbiotic microbiota in the complex processes of lignocellulose degradation and conversion. Especially conspicuous is our poor understanding of the hindgut communities of wood-feeding 'higher' termites, the most species-rich and abundant of all termite lineages[4]. Higher termites do not contain hindgut flagellate protozoa, which have long been known to be sources of cellulases and hemicellulases in the 'lower' termites. The host tissue of all wood-feeding termites is known to be the source of one cellulase, a single-domain glycohydrolase family 9 enzyme that is secreted and active in the anterior compartments of the gut tract[5]. Only in recent years has research provided support for a role of termite gut bacteria in the production of relevant hydrolytic enzymes. That evidence includes the observed tight attachment of bacteria to wood particles, the antibacterial sensitivity of particle-bound cellulase activity[2], and the discovery of a gene encoding a novel endoxylanase (glycohydrolase family 11) from bacterial DNA harvested from the gut tract of a *Nasutitermes* species[6]. Here, in an effort to learn about gene-centred details relevant to the diverse roles of bacterial symbionts in these successful wood-degrading insects, we initiated a metagenomic analysis of a wood-feeding 'higher' termite hindgut community, performed a proteomic analysis with clarified gut fluid from the same sample, and examined a set of candidate enzymes identified during the course of the study for demonstrable cellulase activity.

A nest of an arboreal species closely related to *Nasutitermes ephratae* and *N. corniger* (Supplementary Fig. 1) was collected near Guápiles, Costa Rica. From worker specimens, luminal contents were sampled specifically from the largest hindgut compartment, the microbe-dense, microlitre-sized region alternatively known as the paunch or the third proctodeal segment (P3; Fig. 1a). In the interest of interpretive clarity, we specifically excluded sampling from and analysis of the microbiota attached to the P3 epithelium and the other distinct microbial communities associated with the other hindgut compartments.

Total community DNA from pooled P3 luminal contents was purified, cloned and sequenced. About 71 million base pairs of Sanger sequence data were generated and assembled. The assembly was highly fragmented, with the largest contiguous fragment being only about 15 kilobases (kb) long, attesting to a complex community that limited the analysis to being largely centred on genes and gene modules. For a better association of phylogenetic markers with key functional genes, 15 fosmids were selected for further analysis after an initial end-sequencing screen. The fosmid data were also used in the training of the sequence composition-based classifier, PhyloPythia[7], resulting in the classification of 9% of all contigs beyond the level of phylum. The high gene-coding density and the fact that 0.03% of contigs were classified as Arthropoda suggest minimal host DNA sequence in the data. Genes associated with activities expected to be present among epithelial bacterial communities, such as aerobic

[1]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [2]Verenium Corporation (formerly Diversa), 4955 Directors Place, San Diego, California 92121, USA. [3]IBM Thomas J. Watson Research Center, PO Box 218, Yorktown Heights, New York 10598, USA. [4]Oak Ridge National Laboratory, Biosciences Division, Oak Ridge, Tennessee 37831-6026, USA. [5]Department of Environmental Science and Engineering, [6]Division of Biology, Mailcode 138-78, California Institute of Technology, Pasadena, California 91125, USA. [7]INBio, Instituto Nacional de Biodiversidad, Apdo. Postal 22-3100 Santo Domingo de Heredia, Costa Rica. †Present addresses: Synthetic Genomics, Inc., 11149 North Torrey Pines Road, Suite 100, La Jolla, California 92037, USA (T.H.R., E.J.M.); Max Planck Institute for Computer Science, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany (A.C.M.).
*These authors contributed equally to this work.

respiratory metabolisms[8], were rare or absent. From this we conclude that the source of the nucleic acids and proteins analysed in this study is the P3 luminal microbiota.

A PCR-based survey was used to assess the community structure of the P3 lumen. No amplification of 16S rRNA genes was observed with the use of Archaea-specific primers. An analysis of about 1,750 bacterial 16S rRNA gene sequences amplified from the community DNA revealed a broad diversity of bacteria representing 12 phyla and 216 phylotypes (99% sequence identity threshold) (Fig. 1b). Non-parametric diversity estimates indicated that about 80% of the total species diversity had been sampled (Supplementary Table 1 and Supplementary Fig. 2). The ten most frequently recovered phylotypes comprised 47% of the collection. The 24 most frequently recovered phylotypes in the PCR inventory belonged to the genus *Treponema* or the phylum Fibrobacteres (Fig. 1b). It was expected that members of these two phyla would constitute the majority of the source community genomic DNA analysed. Inspection of the metagenomic data for conserved phylogenetic marker genes and PhyloPythia binning of the data set supported this prediction. Treponemes dominated, contributing 68% of marker genes in the metagenomic DNA (Supplementary Table 2). Fibrobacters contributed 13% of the identifiable fragments in the metagenomic data set.

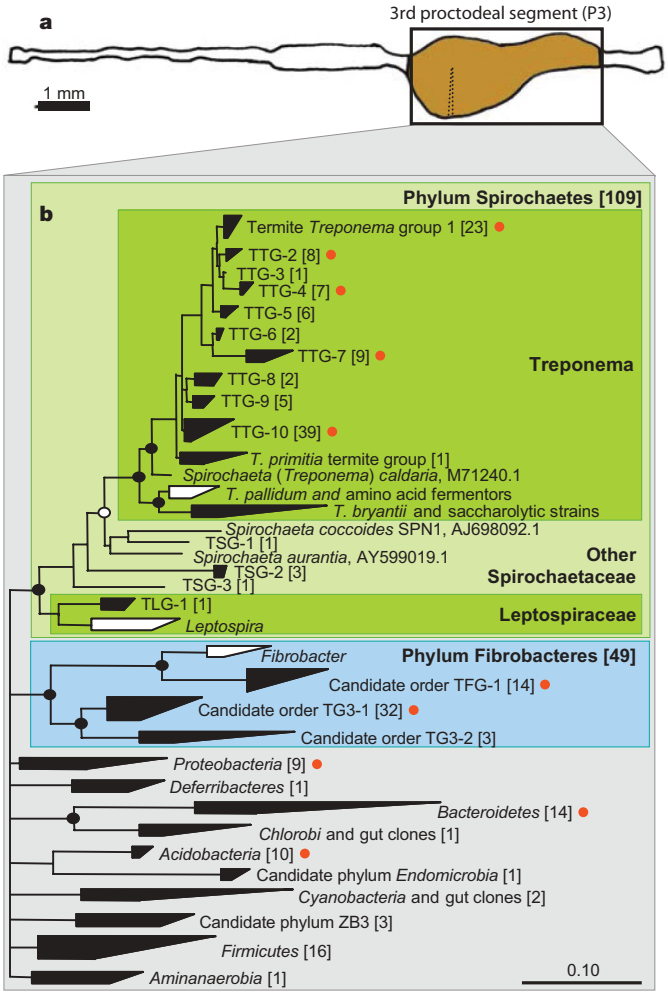

Figure 1 | Source and composition of the bacterial community analysed. **a,** Diagram of the gut tract extracted from the *Nasutitermes* host insect. The anterior of the P3 paunch was incised as indicated (dotted line) with a syringe needle. The luminal contents from about 165 specimens were collected and pooled for nucleic acid and protein analyses. **b**, Phylogenetic diversity of the P3 luminal microbiota. From a PCR-based inventory and from the metagenome libraries, 1,703 almost full-length and 41 full-to-partial 16S rRNA gene sequences were used in a maximum-likelihood analysis (RAxML). The number of distinct P3 luminal community phylotypes contained within each grouping is given in brackets (a total of 216 operational taxonomical units sharing at least 99% sequence identity); black shading denotes that at least one of these phylotypes was represented within the PCR inventory or metagenome libraries (red dots). White shading denotes reference groups having no representation in this or other public inventories derived from termite gut. The phylogram was constructed from 1,289 unambiguously aligned and filtered nucleotide positions. Branching pattern confidence is visualized as follows: no circle, more than 50%; open circle, more than 70%; filled circle, more than 90%. Scale bar indicates 10% estimated sequence divergence. See Supplementary Tables 1, 2 and 11 and Supplementary Figs 2–9 for accession numbers and detailed phylogenetic analyses.

Lignocellulose degradation requires a broad array of enzymes and associated proteins[9]. Consistent with past studies[10], analysis of the *Nasutitermes* P3 luminal community metagenome revealed no evidence for lignin degradation. In contrast, a conservative analysis, detecting complete domains by global alignment (see Supplementary Information), identified many genes and gene modules homologous with more than 700 glycoside hydrolase (GH) catalytic domains corresponding to 45 different CAZy families (carbohydrate-active enzymes; http://www.cazy.org), including a rich diversity of putative cellulases and hemicellulases (Table 1 and Supplementary Tables 3–6).

More than 100 gene modules relevant to cellulose hydrolysis were identified in the data set, corresponding to the catalytic domains of GH5 cellulases, GH94 cellobiose or cellodextrin phosphorylases, GH51 endoglucanase/arabinofuranosidases, and a smaller number of GH8, GH9, GH44, GH45 and GH74 endoglucanases. In contrast, gene modules corresponding to the catalytic domains of GH6 and GH48 family endoglucanases and cellobiohydrolases, key components of several well-studied microbial cellulase systems, were absent.

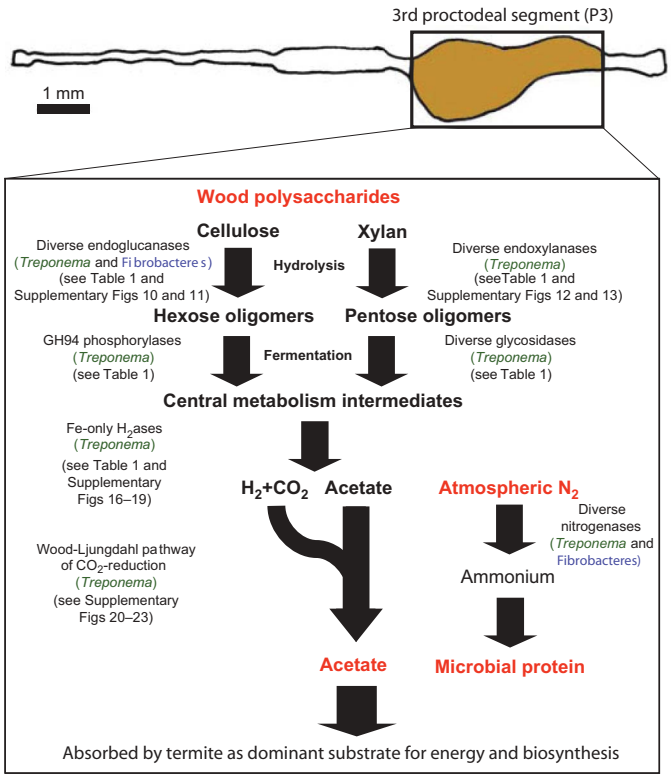

Figure 2 | Model of nutritional symbiosis-relevant metabolism by *Nasutitermes* P3 luminal bacteria. Wood is triturated by the insect's mandibles into small particles and predigested by poorly studied upstream processes before transit into the P3 compartment. The P3 lumen is dominated by diverse species of *Treponema* (Spirochaetes) and Fibrobacteres. There was no evidence for methanogenesis or lignin degradation in the metagenomic data set.

Table 1 | Glycoside hydrolases and carbohydrate-binding modules

CAZy family*	Pfam HMM name†	Known activities‡	Termite gut community§	Source organisms‖	Proteomics¶	Activity#	Signal peptides⋆
Glycoside hydrolase catalytic domains**††							
GH1	Glyco_hydro_1	β-Glucosidase, β-galactosidase, β-mannosidase, others	22	5 trep	1		
GH2	Glyco_hydro_2_C	β-Galactosidase, β-mannosidase, others	23	4 trep			1
GH3	Glyco_hydro_3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-L-arabinofuranosidase, others	69	14 trep	1		22
GH4	Glyco_hydro_4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	14	3 trep	2		2
GH5	Cellulase	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	56	6 trep	2	26 of 30	14
GH8	Glyco_hydro_8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	5				3
GH9	Glyco_hydro_9	Endoglucanase, cellobiohydrolase, β-glucosidase	9	2 fibr		7 of 12	3
GH10	Glyco_hydro_10	Xylanase, β-1,3-endoxylanase	46	11 trep	1		8
GH11	Glyco_hydro_11	Xylanase	14	4 trep			
GH13	Alpha-amylase	α-Amylase, catalytic domain, and related enzymes	48	6 trep	2		
GH16	Glyco_hydro_16	β-1,3(4)-Endoglucanase, others	1				
GH18	Glyco_hydro_18	Chitinase, endo-β-*N*-acetylglucosaminidase, non-catalytic proteins	17	4 trep			6
GH20	Glyco_hydro_20	β-Hexosaminidase, lacto-*N*-biosidase	15	1 trep			
GH23	SLT	G-type lysozyme, peptidoglycan lytic transglycosylase	52	11 trep			9
GH25	Glyco_hydro_25	Lysozyme	1	1 trep			
GH26	Glyco_hydro_26	β-1,3-Xylanase, mannanase	15	3 trep			3
GH27	Melibiase	α-Galactosidase, α-*N*-acetylgalactosaminidase, isomalto-dextranase	4				
GH28	Glyco_hydro_28	Polygalacturonase, rhamnogalacturonase, others	6	1 trep			
GH31	Glyco_hydro_31	α-Glucosidase, α-xylosidase, others	26	7 trep			1
GH35	Glyco_hydro_35	β-Galactosidase	3	1 trep			
GH36	No Pfam§	α-Galactosidase, α−*N*-acetylgalactosaminidase	5				
GH38	Glyco_hydro_38	α-Mannosidase	11	3 trep			1
GH39	Glyco_hydro_39	β-Xylosidase, α-L-iduronidase	3	2 trep			1
GH42	Glyco_hydro_42	β-Galactosidase	24	4 trep	1		
GH43	Glyco_hydro_43	Xylanase, β-xylosidase, α-L-arabinofuranosidase, arabinanase, others	16	6 trep			1
GH44	No Pfam§	Endoglucanase, xyloglucanase	6	2 trep			
GH45	Glyco_hydro_45	Endoglucanase (mainly eukaryotic, 2 bacterial)	4			2 of 2	2
GH51	No Pfam§	Endoglucanase, α-L-arabinofuranosidase	18	1 arch			
GH52	Glyco_hydro_52	β-Xylosidase	3				
GH53	Glyco_hydro_53	β-1,4-Endogalactanase	12	2 trep			
GH57	Glyco_hydro_57	α-Amylase, 4-α-glucanotransferase, α-galactosidase, others	17	6 trep			
GH58	No Pfam§	Endo-*N*-acetylneuraminidase or endo-sialidase	1				
GH65	Glyco_hydro_65m	Trehalase, maltose phosphorylase, trehalose phosphorylase	6				
GH67	Glyco_hydro_67M	α-Glucuronidase, others	10	4 trep			
GH74	No Pfam§	Endoglucanase, cellobiohydrolase, xyloglucanase	7				2
GH77	Glyco_hydro_77	4-α-Glucanotransferase, amylomaltase	14	4 trep			
GH88	Glyco_hydro_88	D-4,5 Unsaturated β-glucuronyl hydrolase	9	3 trep	1		
GH91	No Pfam§	Inulin fructotransferase	1				
GH92	Glyco_hydro_92	α-1,2-Mannosidase	2				
GH94	No Pfam§	Cellobiose phosphorylase, chitobiose phosphorylase, cellodextrin phosphorylase	68				
GH95	No Pfam§	α-L-Fucosidase	12				
GH98	Glyco_hydro_98M	Endo-β-galactosidase	1				
GH103	TIGR: MltB	Peptidoglycan lytic transglycosylase	3				
GH106	No Pfam§	α-L-Rhamnosidase	2				
GH109	No Pfam§	α-*N*-Acetylgalactosaminidase	3				

A similar paucity of GH6 and GH48 modules has been observed in the genomes of other cellulose-degrading bacteria[11,12]. A phylogenetic analysis of the GH5 cellulase diversity revealed that most constituted nine unique subclusters (Supplementary Fig. 10), none of which affiliated with *Clostridium* cellulosome GH5 cellulases. Functional genomic screens confirm that at least one example from each of five major, novel GH5 clades showed activity on acid-solubilized and microcrystalline cellulose (Table 1). At least 14 proteins with a GH5 module have an identifiable signal peptide, and metaproteomic analysis suggests that at least several are secreted into the P3 luminal fluid (Table 1). The metagenomic data suggest that the final metabolism of the oligosaccharides into simple sugars proceeds through the activity of GH3 glucosidase modules (at least 23 have an identifiable signal peptide), or through phosphorolytic reactions catalysed by a diversity of GH94 cellobiose and cellodextrin phosphorylases after substrate transport into the bacterial cytoplasm.

About 100 gene modules corresponding to the catalytic domains of GH10, GH11, GH26 and GH43 hemicellulases were also identified (Table 1 and Supplementary Figs 12 and 13). The metaproteomic analysis revealed that at least one of the diverse GH10 endoxylanases is expressed in the clarified P3 luminal fluid (Table 1); eight encode identifiable signal peptides. None of the xylanase modules have yet been examined for xylanase activity, but several cluster phylogenetically with a previously identified GH11 xylanase (Supplementary Fig. 13). Catalytic modules corresponding to several potential pentosidase and hemicellulose-debranching enzyme domains were also identified. In addition to the large number of GH3 modules noted above, many modules corresponding to the catalytic domains of GH1, GH2, GH4, GH67 and GH95 pentosidases were identified. Additionally, between 4 and 34 gene modules corresponding to the catalytic domains of carbohydrate esterase families 2, 4 and 6 acetyl xylan esterases were identified. Only a few gene modules for catalytic domains of polysaccharide lyases (PL) were observed (five PL1 and five PL11). On the basis of marker gene and composition-based binning analyses, we predict that most of the retrieved glycoside hydrolase gene modules are encoded by treponemes (Fig. 2 and Table 1). A notable exception to this is that the catalytic modules of GH45 and GH9 seemed to be encoded solely by fibrobacters.

Fewer putative polysaccharide-binding domains (carbohydrate-binding modules; CBMs) representing only five CAZy families (http://www.cazy.org) were identified in the data set. Modules of

Table 1 | Continued

CAZy family*	Pfam HMM name†	Known activities‡	Termite gut community§	Source organisms‖	Proteomics¶	Activity#	Signal peptides★
Carbohydrate-binding domains‡‡§§							
CBM4	CBM_4_9	Amorphous cellulose-, xylan- and glucan-binding domain	5		1		1
CBM6	CBM_6	Amorphous cellulose- and xylan-binding domain	13	2 trep			1
CBM11	CBM_11	Glucan-binding domain	3				
CBM30	No Pfam§	Cellulose-binding domain	1				
CBM32	F5_F8_type_C	Galactose- and lactose-binding domain	4	1 fibr			
CBM36	No Pfam§	Xylan-binding domain	2				
CBM37	No Pfam§	Broad binding specificity	1				
Other domains often associated with GH catalytic domains‖ ‖							
	Alpha-L-AF_C	α-L-Arabinofuranosidase, C-terminal domain; associated with GH51	10	1 trep			
	Alpha-mann_mid	Middle domain; associated with GH38	15	3 trep			
	Big_2	Bacterial Ig-like domain, group 2	140	7 trep			31
	Big_3	Bacterial Ig-like domain, group 3	22				
	Bgal_small_N	β-Galactosidase, small chain; associated with GH2	5				
	CBM_X	Associated with GH94	44	8 trep	8		
	CelD_N	N-terminal Ig-like domain of cellulase; associated with GH9	5	1 fibr			4
	fn3	Fibronectin type III domain	45	4 trep			3
	GDE_C	Amylo-α-1,6-glucosidase	1				
	Glyco_hydro_2	Immunoglobulin-like β-sandwich domain	4				
	Glyco_hydro_2_N	Sugar-binding domain	26	5 trep, 2 fibr			
	Glyco_hydro_3_C	C-terminal domain (glycan-binding?)	42	8 trep, 1 fibr	3		
	Glyco_hydro_38C	C-terminal domain	12	2 trep			
	Glyco_hydro_42C	C-terminal domain	8	1 trep			
	Glyco_hydro_42M	Trimerization domain	20	4 trep			
	Glyco_hydro_65C	C-terminal domain	3				
	Glyco_hydro_65N	N-terminal domain	1				
	Glyco_hydro_67C	C-terminal domain	11	2 trep			
	Glyco_hydro_67N	N-terminal domain	2				
	Glyco_transf_36	Associated with GH94	47	12 trep	8		
	GT36_AF	Associated with GH94	45	13 trep	8		
	He_PIG	Putative Ig-like domain	11	1 trep			
	Isoamylase_N	Isoamylase N-terminal domain, associated with GH13	15	2 trep			1

For a complete inventory and genome comparisons, see Supplementary Tables 3–6.
* CAZy: carbohydrate-active enzymes.
† Pfam, http://www.sanger.ac.uk/Software/Pfam/; HMM, hidden Markov model.
‡ http://www.CAZy.org.
§ Number of Pfam HMMs in the metagenomic data set with e-values smaller than 10^{-4}. For modules with no available Pfam HMM, this corresponds to the number of BLAST hits with e-values smaller than 10^{-6}. See Supplementary Information for representative 'non-Pfam' modules used in BLAST searches.
‖ As determined by sequence composition-based classification using Phylopythia; see the text and Methods for details. (trep, treponeme; fibr, fibrobacter; arch, archaeal)
¶ Number of modules expressed as protein *in situ*, as demonstrated by a metaproteomic analysis of the clarified gut fluid.
# Number of cloned modules having cellulase activity on PASC, out of the number examined for this activity.
★ As determined using SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/); see Methods for details.
** CAZy GH families not represented in the data set: 6, 7, 12, 14, 15, 17, 19, 22, 24, 29, 30, 32, 33, 34, 37, 46, 47, 48, 49, 50, 54, 55, 56, 59, 61, 62, 63, 64, 66, 68, 70, 71, 72, 73, 75, 76, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 89, 90, 93, 96, 97, 99, 100, 101, 102, 107, 108 and 110.
†† CAZy GH families 21, 40, 41, 60, 69, 104 and 105 no longer exist, have been reassigned, or are recognized by the Pfam HMMs of other GH families.
‡‡ CAZy cellulose-binding domains (CBMs) not represented in the data set: 1, 2, 3, 5, 8, 10, 13, 14, 15, 17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 29, 31, 33, 34, 35, 38, 39, 40, 41, 42, 43, 44, 45, 46, 48 and 49.
§§ CAZy CBMs 7, 9, 12, 16, 22, 28 and 47 either no longer exist or are recognized by the Pfam HMMs of other CBM families.
‖ ‖ Several gene modules that are often found associated with glycoside hydrolases that were not represented in the data set include Cohesin, Dockerin_1, Alpha-amylase_C, CHB_HEX, CHB_HEX_C, ChiC, ChitinaseA_N, Glucodextran_B, Glucodextran_N, Glyco_hydro_20b, Glyco_hydro_32C, Glyco_hydro_98C and TIG.

undetermined function often found associated with various glycoside hydrolases were also identified (Table 1). *Clostridium*-like cellulosome domains, such as dockerins and cohesins, were not identified.

A total of 34 groups of genes of unknown function, termite orthologous groups, were found abundant in comparison with available metagenomes from soil[13], seawater[14] and human gut[15] (Supplementary Table 10). The expression of several of these in clarified lumen fluid was verified by metaproteomic analysis (Supplementary Table 10). One uncharacterized, putative extracytoplasmic structural protein domain, TIGR02145, was strikingly overrepresented; it appeared 482 times in the termite gut metagenome (Supplementary Fig. 14). This module is found dozens of times in the genome of the rumen cellulolytic bacterium *Fibrobacter succinogenes* (http://www.tigr.org/tdb/rumenomics)[16] but is otherwise restricted to only a few other microbes. It becomes tempting to speculate that this and several of the other classes of unknown domains highlighted in this study may be involved in cell-surface-associated enzymatic, binding and other functions relevant to lignocellulose conversion. However, this possibility has not yet been examined experimentally.

Free H_2 is generated as a product of the fermentation of cellulose and xylan by several protozoa from lower termites[17], but its production rate, concentration and biological sources in protozoan-free, wood-fedding higher-termites are not yet known. Gene modules coding for 159 diverse iron-only hydrogenases[18] were identified in the metagenomic data set. Many were binned to the genus *Treponema* by PhyloPythia. By examining sequence similarity, domain structure and gene neighbourhoods, we predict that about half of these genes encode catalytic enzymes. Although some of the catalytic hydrogenases are closely related to known enzymes (Supplementary Fig. 15), several were inferred to have unusual spatial arrangements of the cysteines that coordinate the catalytic iron–sulphur clusters (Supplementary Figs 17 and 18). The rest seem to be involved in signal transduction and chemotaxis, as demonstrated by hydrogenases that include for example methyl-accepting chemotaxis domains (Supplementary Fig. 16).

The major electron sink reaction occurring during the fermentation of wood in termites is CO_2-reductive acetogenesis[19]. One-quarter of the gut acetate that ultimately serves as the major carbon and energy source for the insect host can be derived from CO_2 (ref. 20). Little is known about the diversity of CO_2-reducing acetogens in the gut tracts of higher, wood-feeding termites. Analysis of the data set identified 14–37 variants of all except one (formate dehydrogenase) of the known proteins associated with the Wood–Ljungdahl pathway[21] of CO_2-reductive acetogenesis (Supplementary Fig. 19). Phylogenetic analysis of a marker gene of the pathway, that encoding formyl-tetrahydrofolate synthetase (FTHFS), revealed that most of the metagenome-derived genes clustered with orthologues from the genuine homoacetogen *Treponema primitia*

(Supplementary Fig. 21) and with other FTHFS genes that occur and are expressed in the guts of lower termites[22]. A key multi-enzyme complex of the pathway is carbon monoxide dehydrogenase (CODH) acetyl-CoA synthase[21]. Analysis of variants of hindgut CooS (the gene encoding the CODH catalytic subunit) revealed two distinct and novel phylogenetic clusters (Supplementary Fig. 22), both predicted to be encoded by treponemes by PhyloPythia. In the higher termite *Nasutitermes*, the symbiotic process of CO_2-reductive acetogenesis seems to be dominated by spirochetes, perhaps to the near or total exclusion of *Sporomusa termitida* or other spore-forming Firmicutes, historically the better-studied representatives of this physiology.

Genes encoding formate dehydrogenases (and related molybdopterin enzymes in general) and ion-translocating hydrogenases typically associated with the pathway were nearly absent from the data set. Formate is an obligatory pathway intermediate (Supplementary Figs 19 and 20), so the near absence of FDH remains unexplained. The scarcity of ion-translocating hydrogenases might be compensated for by the relative enrichment of RnfC-like ion-translocating NADH:ferredoxin oxidoreductases and ferredoxin-domain membrane proteins of unknown function (Supplementary Fig. 19).

Wood is depleted in nitrogen and is therefore a poor source of essential amino acids and protein. Bacterial N_2 fixation has been shown to affect the biology of Costa Rican *Nasutitermes* and other termite species[23], and a rich diversity of *nif*H genes has been inventoried in the hindguts of a *Nasutitermes* species from Japan[24]. Twelve near-full-length *nif*H homologues were identified in this analysis of the P3 luminal community. Composition-based predictions did not successfully bin any of the P3 NifH homologues to known genera or phyla. Between 31 and 100 homologues of other nitrogenase components (NifD, NifK, NifE and NifN) were identified in the data set; at least one variant of each identified could be binned to the treponemes or to the fibrobacters, implicating species from both of these phyla in N_2 fixation in the gut (Fig. 2). The discrepancy between the contrasting recoveries of NifH and, for example, NifE homologues remains unexplained. The host's nitrogen waste, uric acid, has previously been shown to be fermented and recycled by *Streptococcus* and *Sebaldella* species resident in certain termite hindguts[25]. However, there is as yet insufficient information on uric acid fermentation genes to be able to examine the data set for their presence.

Termite guts are characterized by steep physical, chemical and biological gradients[26]. Human colonic and bovine ruminal microbial communities are dominated by non-motile microbes, but termite hindguts have long been noted as sources of highly motile bacteria[27]. Motility in bacteria is often coupled with chemotactic behaviour. Indeed, about 1,500 genes related to chemotaxis and chemosensation were identified in the data set (Supplementary Fig. 23). The abundance of these genes in the termite gut suggests that chemotaxis may be relevant both during the colonization, by means of trophallaxis, of the microbe-free gut in young termites and after each moult, and also after regular peristaltic mixing and redistribution of the gut contents, thus maintaining compartmentalization and promoting niche selection.

This study illustrates how complex a 1-μl environment can be. The *Nasutitermes* species examined here and other termites may be considered to be rich reservoirs of bacterial enzymes relevant both to reaching a better understanding of the fundamental process of wood degradation, and to engineering novel schemes for the conversion of lignocellulose into biofuels and other microbial products of interest.

METHODS SUMMARY

Termites collected under permit near Guápiles, Costa Rica, were classified morphologically and by cytochrome oxidase gene sequence (Supplementary Fig. 1). Dissections were performed within 36 h of collection. Hemi-transverse incisions were made to the anterior P3 hindgut compartment. Luminal contents from about 165 worker specimens were pooled, diluted into buffered saline, and frozen immediately. Whole-genome shotgun libraries containing inserts were prepared and sequenced. Data were assembled using Phrap; coding regions were predicted using fgenesb. Fosmids were nearly completely sequenced using pyrosequencing and traditional methods. Fragments 1 kb or larger were classified using PhyloPythia. Assembled data were incorporated into the Integrated Microbial Genomes with Microbiome Samples (IMG/M) system[32] (http://img.jgi.doe.gov/m). For protein analysis, aliquots of the luminal contents in buffered saline were clarified by centrifugation. Protein from the soluble fraction was heat-treated, reduced, alkylated and digested. Digested samples were analysed by means of three-dimensional LC–MS/MS with XCalibur Rawfile Converter V, and searched against the termite metagenome database with SEQUEST. Searches for glycoside hydrolases and carbohydrate-binding modules were performed using HMMER hmmsearch with Pfam hidden Markov models (HMMs). GHs and CBMs were named using CAZy (carbohydrate-active enzymes) nomenclature. When a Pfam HMM for a given GH either did not exist or did not correspond to the catalytic module, BLAST searches against the data set were performed using regions of sequences listed at the CAZy website. Assays for activity were conducted with phosphoric-acid-swollen cellulose (PASC) and microcrystalline cellulose. 16S rRNA genes were amplified using universal primers, sequenced, and analysed with ARB and PHYLIP. High-quality sequence reads are deposited in the NCBI trace archive.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 July; accepted 18 September 2007.

1. Sugimoto, A., Bignell, D. E. & Macdonald, J. A. in *Termites: Evolution, Sociality, Symbioses, Ecology* (eds Abe, T. Bignell, D. E. & Higashi, M.) 409–435 (Kluwer Academic, Dordrecht, 2000).
2. Tokuda, G. & Watanabe, H. Hidden cellulases in termites: revision of an old hypothesis. *Biol. Lett.* **3,** 336–339 (2007).
3. Brune, A. in *The Prokaryotes* Vol. 1 (eds Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E.) 439–474 (Springer, New York, 2006).
4. Kambhampati, S. & Eggleton, P. in *Termites: Evolution, Sociality, Symbioses, Ecology* (eds Abe, T. Bignell, D. E. & Higashi, M.) 1–24 (Kluwer Academic, Dordrecht, 2000).
5. Tokuda, G. *et al.* Major alteration of the expression site of endogenous cellulases in members of an apical termite lineage. *Mol. Ecol.* **13,** 3219–3228 (2004).
6. Brennan, Y. *et al.* Unusual microbial xylanases from insect guts. *Appl. Environ. Microbiol.* **70,** 3609–3617 (2004).
7. McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* **4,** 63–72 (2007).
8. Brune, A., Emerson, D. & Breznak, J. A. The termite gut microflora as an oxygen sink: microelectrode determination of oxygen and pH gradients in guts of lower and higher termites. *Appl. Environ. Microbiol.* **61,** 2681–2687 (1995).
9. Davies, G. J. & Henrissat, B. Structural enzymology of carbohydrate-active enzymes: implications for the post-genomic era. *Biochem. Soc. Trans.* **30,** 291–297 (2002).
10. Breznak, J. A. & Brune, A. Role of microorganisms in the digestion of lignocellulose in termites. *Annu. Rev. Entomol.* **39,** 453–487 (1994).
11. Taylor, L. E. II *et al.* Complete cellulase system in the marine bacterium *Saccharophagus degradans* strain 2-40T. *J. Bacteriol.* **188,** 3849–3861 (2006).
12. Xie, G. *et al.* Genome sequence of the cellulolytic gliding bacterium *Cytophaga hutchinsonii*. *Appl. Environ. Microbiol.* **73,** 3536–3546 (2007).
13. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308,** 554–557 (2005).
14. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
15. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
16. Qi, M. *et al.* Novel molecular features of the fibrolytic intestinal bacterium *Fibrobacter intestinalis* not shared with *Fibrobacter succinogenes* as determined by suppressive subtractive hybridization. *J. Bacteriol.* **187,** 3739–3751 (2005).
17. Odelson, D. A. & Breznak, J. A. Nutrition and growth characteristics of *Trichomitopsis termopsidis*, a cellulolytic protozoan from termites. *Appl. Environ. Microbiol.* **49,** 614–621 (1985).
18. Vignais, P. M. & Colbeau, A. Molecular biology of microbial hydrogenases. *Curr. Issues Mol. Biol.* **6,** 159–188 (2004).
19. Breznak, J. A. & Switzer, J. M. Acetate synthesis from H2 plus CO2 by termite gut microbes. *Appl. Environ. Microbiol.* **52,** 623–630 (1986).
20. Pester, M. & Brune, A. Hydrogen is the central free intermediate during lignocellulose degradation by termite gut symbionts. *ISME J.* **1,** 551–565 (2007).
21. Ragsdale, S. W. The eastern and western branches of the Wood/Ljungdahl pathway: how the East and West were won. *Biofactors* **6,** 3–11 (1997).
22. Pester, M. & Brune, A. Expression profiles of fhs (FTHFS) genes support the hypothesis that spirochaetes dominate reductive acetogenesis in the hindgut of lower termites. *Environ. Microbiol.* **8,** 1261–1270 (2006).

23. Prestwich, G. D., Bentley, B. L. & Carpenter, E. J. Nitrogen sources for neotropical nasute termites: Fixation and selective foraging. *Oecologia* **46,** 397–401 (1980).
24. Ohkuma, M., Noda, S. & Kudo, T. Phylogenetic diversity of nitrogen fixation genes in the symbiotic microbial community in the gut of diverse termites. *Appl. Environ. Microbiol.* **65,** 4926–4934 (1999).
25. Potrikus, C. J. & Breznak, J. A. Anaerobic degradation of uric acid by gut bacteria of termites. *Appl. Environ. Microbiol.* **40,** 125–132 (1980).
26. Brune, A. & Friedrich, M. Microecology of the termite gut: structure and function on a microscale. *Curr. Opin. Microbiol.* **3,** 263–269 (2000).
27. Beckwith, T. & Light, S. The spirals within the termite gut for class use. *Science* **66,** 656–657 (1927).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank production and sequencing teams at Verenium and the Joint Genome Institute for their expertise, J. Mata for morphological identifications, L. Christoffersen for logistic and permitting support, and our laboratory colleagues for their comments during manuscript preparation. This research was supported in part by the National Science Foundation, the National Institutes of Health, Caltech, and the Lawrence Berkeley National Laboratory. The sequencing for the project was provided through the US Department of Energy (DOE) Community Sequencing Program (http://www.jgi.doe.gov/CSP/index.html). This work was performed, in part, under the auspices of the DOE Office of Science, Biological and Environmental Research Program, University of California, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory.

Author Contributions F.W., P.H., E.J.M., D.E.R., E.M.R. and J.R.L. performed project planning, coordination, execution and facilitation. M.H., C.M., L.G.A. and G.T. undertook field research planning, permits, logistics and station management. F.W., M.C., M.H., C.M., L.G.A. and J.R.L. conducted field collection and sample preparation. M.C., G.D., N.A., J.M. and C.C. performed nucleic acid purification and library construction. D.P. and K.B. carried out assemblies. A.S. conducted gene prediction and annotation. E.S. undertook data processing and loading into IMG/M. N.I. and N.C.K. performed metabolic reconstruction. A.C.M. and I.R. carried out binning. N.M. conducted fosmid annotation and manual curation. M.G., J.T.S. and B.D.G. performed proteomics and enzyme activities. F.W., P.L., V.K., D.D., E.K., E.G.M., E.A.O. and X.Z. carried out phylogenetic analyses. H.G.M. made accumulation curves, diversity estimates, statistical test for gene-centric analysis. P.L., T.H.R. and J.R.L. performed glycoside hydrolase bioinformatics. R.S. and S.G.T. constructed hypothetical gene families. N.I. and P.H. were responsible for hydrogenases. E.G.M., E.A.O., X.Z. and J.R.L. performed C1-pathway and N-metabolic reconstruction. M.P. carried out sensory transduction protein analysis. F.W., P.L., M.G., T.H.R., J.T.S., P.H., N.I., R.S., S.G.T., M.P. and J.R.L. undertook manuscript preparation.

Author Information This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession number ABDH00000000; this first version is ABDH01000000. The COII gene from the termite host is deposited under accession number EU236539, and 16S rRNA gene sequences are deposited under the accession numbers EF453758–EF455009 and EU024891–EU024927. The subcloned cellulase gene sequences are deposited under the accession numbers EF428062–EF428109. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.R.L. (jleadbetter@caltech.edu).

METHODS

Sample collection. Termite collection '290cost002' was made during the late morning of 24 May 2005 within the secondary forest of Bosque Lluvioso, an INBio private reserve near the town of Guápiles, Costa Rica (latitude 10° 11′ 26.0″ N, longitude 83° 51′ 34.5″ W). The nest was about 0.5 m in diameter and was attached to the trunk of an *Alchorneopsis floribunda* tree at a height of about 1 m. Specimens from this and other nests in the nearby area were observed to feed preferentially on dead limbs of *Cecropia obtusifolia*. At the time of collection, the temperature within the nest was 28 °C. The nest was cut into pieces with a machete and returned to InBio's laboratory in Heredia, where all dissections were performed within the first 36 h of collection. Specimens were identified morphologically as being most similar to *Nasutitermes ephratae*.

DNA extraction. For bacterial DNA and luminal protein analysis, gut tracts of ice-chilled, robust worker specimens were extracted using clean, sharp fine-tipped forceps, and arrayed on a piece of clean Parafilm. A hemi-transverse incision of the P3 hindgut compartment was made with the tip of a sterile 23-gauge needle attached to a 1-ml tuberculin syringe. Immediately thereafter, 7 µl of an anoxic buffered saline solution (BSS; 10.8 mM K_2HPO_4, 6.9 mM KH_2PO_4, 21.5 mM KCl, 24.5 mM NaCl, 0.5 mM $CaCl_2$, 10.0 mM $NaHCO_3$, 5.3 mM $MgCl_2$ pH 7.2) was placed over the incision; the BSS had been filtered through a 0.2-µm pore-size filter into sterile, crimp-sealed tubes containing an atmosphere of 100% N_2. Tubes of BSS were vacuum-purged for six cycles with sterile 100% N_2 to lower the dissolved O_2 concentration. The tip of a sterile, 200-µl plastic pipette tip attached to a standard micropipetter was used to press the dense, tan-coloured luminal contents gently out of the P3 compartment. The contents were stirred into the BSS, aspirated into the pipette and pooled into a microcentrifuge tube maintained on ice, and frozen at −80 °C immediately thereafter. The pooled P3 luminal contents from 165 worker specimens were used in all subsequent analyses of P3 luminal contents. DNA was isolated from the pooled luminal contents as described previously[28]. DNA was also extracted from heads of soldier termites collected from nest 290cost002 with phenol/chloroform to confirm the identity of the host species.

Metagenome processing: shotgun library preparation, sequencing and assembly. Two shotgun libraries from the P3 genomic DNA were prepared: a 2–4-kb insert library cloned into pBK-CMV and a roughly 32-kb insert fosmid library cloned into pCC1Fos (Epicentre Corp.)[29]. Both libraries were sequenced with BigDye Terminators v3.1 and resolved with ABI PRISM 3730 (ABI) sequencers. 92,160 reads comprising 68.47 megabases (Mb) of phred Q20 sequence were generated from the small insert library, and 13,824 reads comprising 2.91 Mb of phred Q20 sequence were generated from the pCC1Fos library. The reads were base-called using phred version 0.990722.g (refs 30, 31), and megablasted against the NCBI nucleotide database and internal contaminant databases. Reads with significant hits on known contaminants and/or with fewer than 100 Q20 base pairs were eliminated from the data set. The remaining 95,324 reads were assembled using parallel phrap (http://www.phrap.org) compiled for the Sun operating system, version SPS 3.57. The resulting assembly consisted of 41,765 contigs covering 44.4 trimmed consensus megabase pairs. The longest contig was 14.7 kbp long and contained 72 reads. However, 25% of the reads remained as singlets. This draft quality assembly was used for all downstream analysis.

Full fosmid sequencing and assembly. Fifteen termite hindgut fosmids were chosen for full sequencing. The fosmids were grown overnight in Luria–Bertani (LB) broth, normalized for cell density and pooled together with 15 fosmids from different projects for DNA extraction using the Qiagen Genomic-tip 100/G kit in accordance with the manufacturer's instructions. Extracted DNA was directly sequenced with the 454 Life Sciences Genome Sequencer GS20 and about 30 million base pairs of about 100-bp sequence reads were generated from one run. In addition, the pooled fosmid DNA was hydrosheared, subcloned into pUC18 and end-sequenced, producing 7,680 Sanger reads. Sanger and GS20 sequences were co-assembled with the Forge assembler developed at JGI (D.P., unpublished observations). Seven complete and eight partly assembled termite hindgut fosmids, comprising 37 genomic fragments, were assembled and manually verified from the pooled data set.

Gene prediction. The termite hindgut metagenome and fosmid assemblies were annotated with the *ab initio* gene-calling program fgenesb (SoftBerry). Normally *ab initio* gene calling of isolate genomes trains on the data set being annotated; however, because metagenomes are multigenomic data sets, self-training generates low-quality results (data not shown). Instead, parameters were obtained from training on multiple bacterial isolate genomes to provide an 'average' bacterial coding preference and other sequence features such as Shine–Dalgarno sequences. The command string used was bactg_ann.pl mixr_paths_newcog.list1 < sequence_file > 60, where the sequence file is the fasta output of the assembled contigs (for example, in Phrap this is the fasta.screen.contigs file), and 60 is the minimal length of predicted ORFs in base pairs, and mixr_paths_newcog.list1 is a config file that contains information about used programs, databases, etc. This file contains reference to gener.par, which provides generalized 'bacterial' gene parameters. The annotation of the 37 fosmid contigs was curated manually.

Binning. Metagenomic fragments were binned (classified) using PhyloPythia, a phylogenetic classifier that uses a multi-class Support Vector machine (SVM) for the composition-based assignment of fragments at different taxonomic ranks[7]. Generic models for the ranks of domain, phylum and class were combined with sample-specific models for the clades *Treponema* and Fibrobacteres. The generic models represent all clades covered by three or more species at the corresponding ranks among the sequenced microbial isolates. The sample-specific models include classes for the dominant sample populations of *Treponema* and Fibrobacteres, as well as a class 'Other'.

The sample-specific models were each trained on sequence data obtained from fully sequenced fosmids identified using phylogenetic marker genes (see Phylogenetic analysis of conserved single-copy protein-coding genes, below) and in the case of the *Treponema*-specific model, also from fragments of two sequenced *Treponema* genomes (*T. denticola* ATCC 34505 and *T. pallidum* subsp. *pallidum* str. Nichols). Five sample-specific SVMs 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.

Twelve of 15 sequenced fosmids assigned unambiguously through analysis of phylogenetic marker genes were initially used for the training of sample-specific models (four fosmids assigned to Fibrobacteres, corresponding to 94.5 kb of sequence, and eight fosmids assigned to *Treponema*, corresponding to 183.6 kb of sequence). 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 were used, excluding those of the *Treponema* genomes. Composition-based characterization of the complete fosmid sequences with the combined classifier confirmed the marker-gene-based placement for 13 of the 15 fosmids. On the basis of this extended set (five Fibrobacteres fosmids totalling 115.2 kb of sequence and eight *Treponema* fosmids totalling 203 kb of sequence), the final sample-specific models were computed. The final classifier consisting of the sample-specific and generic clade models was 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.

Loading of data into IMG/M. The annotated metagenomic and fosmid sequences (including PhyloPythia assignments) were loaded as independent data sets into IMG/M[32] (http://www.jgi.doe.gov/m), a data-management and analysis platform for genomic and metagenomic data based on IMG[33]. KEGG pathways present in the metagenome were automatically assigned on the basis of EC numbers in the annotation, and pathways not included in the KEGG database were inferred from comparative analysis with other genomes, on the basis of sequence similarity and gene cluster structure conservation. Conserved domain and motif searches were performed with RPS-BLAST, using an e-value threshold of 10^{-2} without low-complexity masking[34]. Overlaps on the query gene were removed as follows: the most significant hit for a query gene was taken first, then the next most signficant hit that did not overlap with the first, and so on, until all overlaps had been removed. SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) was used to identify signal peptides and the information was added to the records of individual genes in IMG/M[35]. No significant differences in signal peptide calls were observed with the Gram-negative and Gram-positive models. Both were tried because of the multi-species nature of the data set.

Metaproteomic analysis: protein extraction and digest. Aliquots of the P3 luminal contents in buffered saline were centrifuged for 10 min (4 °C) at 13,000 r.p.m. on a benchtop microcentrifuge to remove the insoluble fraction of the fluid. Protein (500 µg) from the soluble fraction was heat-treated at 100 °C for 5 min in the presence of 0.5% RapiGest. The samples were reduced with 1 mM dithiothreitol and alkylated with 0.4 mg ml^{-1} iodoacetamide. After digestion with endoproteinase Lys-C (1/200 of the protein), the protein samples were further digested twice with trypsin (1/100 of the protein). The second batch of trypsin was added 4 h after the addition of the first batch. The RapiGest was removed from the solution by the addition of HCl to a final concentration of 0.05 M. Samples were incubated at 37 °C for 60 min and centrifuged at 13,000 r.p.m. on a benchtop microcentrifuge to remove the insoluble fraction. The pH of the soluble fraction was adjusted to pH 2–3 with ammonium hydroxide.

Three-dimensional LC–MS/MS analysis. The digested termite protein samples were analysed by the three-dimensional LC–MS/MS system as described[36]. The LTQ mass spectrometer was set to divide the full MS scan into three smaller

sections covering a total range of 400–2,000 *m*/*z*. Each of the smaller MS scans was followed by four MS/MS scans of the most intense ions from the preceding MS scan. The typical collision energy for collision-induced dissociation was set to 35% with a 30-ms activation time. Dynamic exclusion was enabled with a repeat count of 1 and a 3-min exclusion duration window.

Data analysis. The LC–MS/MS raw data were extracted with the XCalibur Rawfile Converter V 1.0.0a and then searched against the termite metagenome-predicted open reading frames using the SEQUEST program. The non-specific cleavage rule was designated during the SEQUEST search. Differential modifications of Met oxidation (+16) and Cys alkylation (+57) were allowed for the database search. The results were filtered using the same criteria set as described previously[36] to obtain the peptide identifications. In brief, all peptide identifications had a delta cross-correlation score of more than 0.1; peptides with a +1 charge state were fully tryptic with a cross correlation (X_{corr}) of more than 1.9; peptides with a +2 charge state were partly tryptic with $X_{corr} > 3.0$ or fully tryptic with X_{corr} ranging between 2.2 and 3.0; peptides with a +3 charge state were fully or partially tryptic with $X_{corr} > 3.5$. The Portfolio program was used for protein identification and summarizing the results. For each protein sequence in the database, the program assembled the peptide identifications that matched its substring. The protein identification was established if one or more peptide identifications were matched. The peptide identifications were allowed for multiple protein identifications. The number of distinct peptides and sequence coverage per protein identification were also calculated. Proteins identified by a single peptide are considered at a reduced level of confidence. These protein identifications by single peptide were included in all results.

Microbial community composition analysis and termite host identity: 16S rRNA gene PCR clone libraries. Two clone libraries were prepared from pooled PCR products amplified from the P3 genomic DNA using two different primer pairs broadly targeting the bacterial domain, namely 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′). Attempts to amplify archaeal 16S rRNA genes were unsuccessful. PCR amplicons were cloned into the vector pCR4-TOPO (Invitrogen Corp.) and plated onto Carbenicillin-containing (150 μg ml^{-1}) LB agar plates, and five 384-well microtitre plates were picked. Clones were end-sequenced using vector primers, and the sequence reads were vector trimmed, assembled, quality-checked and chimaera-checked using the genelib software package (E. Kirton, unpublished observations). Ten putative chimaeras were identified and excluded from the data set. A total of 1,703 near-complete bacterial 16S rRNA genes passed the quality and chimaera filters in genelib and were used in the subsequent analyses.

Phylogenetic analysis of 16S rRNA gene sequences. The 1,703 sequences were aligned using the NAST aligner[37] and imported into an ARB database with the same alignment (⟨http://greengenes.lbl.gov/⟩; ref. 38). Forty-one partial and near complete 16S sequences were extracted from the P3 metagenomic data set using the NAST aligner and also imported into ARB[48]. Sequences were initially assigned to phylogenetic groups using the ARB Parsimony insertion tool. *De novo* phylogenetic trees (Fig. 1 and Supplementary Figs 3 and 4) were constructed from masked ARB alignments (to remove ambiguously alignable positions) using RAxML[39]. The bootstrapped tree was calculated using the following settings: rapid hill climb, general time-reversible model and optimization of site-specific evolutionary rates with 100 bootstrap resamplings while we used simulated annealing with the same settings but with the time limit of four times the average run time of the rapid hill-climb calculation for the snapshot tree. Snapshot and bootstrapped RAxML topologies were subsequently reimported in Newick format into ARB for visualization. The phylum-level trees (Supplementary Figs 5–9) were reconstructed using TREE-PUZZLE[40] in ARB.

Accumulation curves and diversity estimates. 16S rRNA gene sequences from the two PCR clone libraries were assigned to clusters (operational taxonomical units; OTUs) at 97%, 98%, 99% and 100% sequence identity thresholds using the DOTUR package[41] applied to a distance matrix generated from an unmasked alignment in ARB (Supplementary Table 1). Accumulation curves were constructed by plotting the number of unique OTUs at a 99% identity threshold found in a given number of clones versus the number of clones (Supplementary Fig. 2). The clone order is arbitrary, so the accumulation curve was averaged over a 1,000 random permutations of the initial clone order. Accumulation curves were also plotted for subsets of the OTUs comprising the major phylogenetic lineages identified in the data set. Diversity estimates were calculated using four different methods involving parametric (exponential[42] and two-parameter hyperbola[42,43]) and non-parametric (bootstrap[44–46] and jackknife[44,45]) estimators (Supplementary Table 1). Neither the exponential nor the two-parameter hyperbola models offered good fits, so they were not considered further. The non-parametric estimates for these data produced more meaningful estimates (errors were much smaller than averages) of the total diversity (Supplementary Table 1). On the basis of these estimates, 71–87% of the total OTU diversity had been sequenced.

Phylogenetic analysis of conserved single-copy protein-coding genes. Orthologues belonging to ten conserved single-copy protein-coding gene families were retrieved from the P3 metagenomic and isolate genome data sets using an export tool available in IMG/M[32]. Genes from the termite data set had to be more than 25% of their full-length orthologue in *Escherichia coli* to be included in the analysis. Multiple sequence alignments were created using Muscle[47] and phylogenetic trees were reconstructed using the ARB software package[48]. If sequences clustered clearly with the available *Treponema* sequences (*T. pallidum* and *T. denticola*) or *Fibrobacter succinogenes*, they were assigned to these phylum-level groupings. If no clear relationship with these phyla could be established they were assigned to 'other' (see Supplementary Table 2). Note that these 'other' sequences could still belong to either the Fibrobacteres or Spirochaeta because the available reference genome sequences are related only distantly to the termite hindgut bacteria.

Identification of termite host species. The mitochondrial cytochrome oxidase II (COII) gene was PCR-amplified from genomic DNA extracted from termite soldier heads using the primers A-tLeu_mod (5′-CAG ATA AGT GCA TTG GAT TT-3′) and B-tLys (5′-GTT TAA GAG ACC AGT ACT TG-3′) and was sequenced without cloning. This gene is commonly used for the identification of termite species[49,50]. Reference COII nucleotide sequences were extracted from the public databases and aligned to the termite sequences in ARB. Nucleotide and amino-acid-based trees were constructed using TREE-PUZZLE in ARB (Supplementary Fig. 1) and the topologies were compared.

Glycoside hydrolases and carbohydrate-binding modules: annotation. Database searches for GHs and CBMs were performed using HMMER hmmsearch with Pfam HMMs. The Pfam_ls HMMs were used to find complete matches to the family by global alignment. All hits with e-values smaller than 10^{-4} were counted and their sequences were further analysed. GHs and CBMs were named in accordance with the CAZy nomenclature scheme[9,51,52]. The data sets analysed comprised the 62-Mb assembly of the termite P3 hindgut metagenome, the human distal gut microbiome (subjects 7 and 8), the soil (Diversa silage) metagenome, the completed genomes of *Saccharophagus degradans* (*Microbulbifer degradans*), *Cytophaga hutchinsonii* and *Thermobifida fusca*, and the draft genomes of *Fibrobacter succinogenes*, *Clostridium thermocellum* and *Caldicellulosiruptor saccharolyticus*. The closest related protein sequences for all metagenome sequences were identified with BLAST searches in GenBank and added to the data set. For several CAZy GH and CBM families, no Pfam HMM is available. In these cases, representative sequences were selected from the CAZy website, and the sequence region corresponding to the family of interest was determined. These sequence regions were then used in BLAST searches against the termite gut community metagenome to identify potential sequences belonging to these CAZy GH and CBM families. An e-value cutoff of 10^{-6} was used in these searches. The sequences of the domains used in these searches are listed in Supplementary Information.

Phylogenetic analysis. Sequence alignments were obtained using HMMER hmmalign to the corresponding Pfam HMM. Regions that did not align to the Pfam HMMs were masked from the final alignments. Phylogenetic analysis was performed with the program PROML, which is part of the PHYLIP package, using a maximum-likelihood method with the Jones–Taylor–Thornton (JTT) amino-acid change model, equal rates, and one random sequence addition followed by global rearrangements.

Subcloning and expression of glycoside hydrolases and activity assay. Subclone sequences that were discovered in the termite gut community DNA library by hybridization or activity screening, or were selected from termite gut community metagenome (data mining) and retrieved in full length, were assayed for activity on PASC[53,54], and were added to the data set as well. Of the 33 active enzymes, 32 were identified through heterologous expression from the pSE420 vector in *E. coli*, the other (GenBank EF428075) was identified through expression in *Pichia pastoris*. Whole-cell lysates were prepared and tested for cellulase activity on a simple PASC substrate (50 mM sodium acetate pH 5 or 50 mM $NaPO_4$ pH 7, 0.75% PASC, 1:25 dilution of cell lysate, 37 °C overnight)[55]. Activity was monitored by glucose oxidase/β-glucosidase enzyme couple reactions for cellobiose and glucose release or with the bicinchoninic acid assay (BCA) for reducing-sugar formation. Cellulose assays for glucose/cellobiose release employed the Invitrogen Amplex Red Glucose/Glucose Oxidase Assay Kit (A22189) or the BCA assay kit (Pierce 23221). Enzymes were considered positive for cellulase activity if they produced a signal in these assays greater than one standard deviation from the average activity observed by a vector-only control lysate.

Hypothetical gene families. For the analysis of overrepresented gene families of unknown function, the STRING database[56] was first used to assign COG (Clusters of Orthologous Groups) or NOG (Non-supervised Orthologous

Group) accession to each protein in the set. COG/NOG groups containing 50 or more members in the termite gut microbiota genes and belonging to functional categories 'S' ('function unknown') or 'R' ('general function prediction only') were further analysed. Proteins that were not assigned any COG/NOG were clustered using an all-versus-all blastp (default parameters) followed by mcl clustering as described ⟨http://micans.org/mcl/⟩. Representation of each of the above-described clusters of homologues was calculated in the soil[13] and seawater[14] metagenomes, and clusters of hypothetical proteins that showed at least threefold overrepresentation in the termite gut microflora sequences were considered as termite orthologous groups (TOGs; Supplementary Table 12).

The statistical significance of gene-family overrepresentation was determined by first assigning a relative representation to each orthologous group. The relative family (COG, NOG or TOG) representation r_{ij} ($i = 1, \ldots, F$ is a unique number given to each COG, NOG and TOG, and $j = 1, \ldots, 4$ is a number given to each of the environments) was calculated by normalizing the total number of ORFs of a family i in an environment j, C_{ij}, by the total number of ORFs in each environment:

$$p_{ij} = C_{ij} / \sum_{i=1}^{F} C_{ij}$$

and then normalizing to unity all environments:

$$r_{ij} = p_{ij} / \sum_{j=1}^{4} p_{ij}$$

The relative family representation pinpoints families that are overrepresented in one data set with respect to the others: for example, a value of $r_{i1} = 0.8$ indicates that 80% of the normalized ORF counts of family i are found in the termite hindgut environment (first environment) and 20% are spread over the remaining environments.

Additionally, it would be desirable to have a measure of the robustness of the family representation value. For example, if for two families k and l the counts were $C_{k(j=1,\ldots,4)} = [1\,0\,0\,0]$ and $C_{l(j=1,\ldots,4)} = [423\,0\,0\,0]$, both family representation values r_{j1} and r_{k1} would equal 1 because there are no counts for other data sets. Nonetheless, the result for the second family would be more reliable because it involves a larger number of ORF counts. It is therefore less likely to have been a consequence of noise or finite sampling effects. We quantify this robustness by using a null model in which the possible effects due to finite sampling effects and noise are modelled in the simplest way: by taking 1% of randomly selected ORF counts and reassigning them to randomly chosen families and environments. This is not meant to be a comprehensive modelling of all sources of error, but rather using a basic null model that tests robustness of estimates by introducing a small amount of noise.

The addition of noise can be formally described as follows: if P_{ijk} denotes the presence of ORF k in family i and environment j ($P_{ijk} = 1$ if ORF k belongs to family i and is present in environment j; $P_{ijk} = 0$ otherwise), for a randomly chosen k belonging to family i and present in environment j $P_{ijk} = 1 \rightarrow 0$ and for randomly chosen l and m $P_{ijk} = 0 \rightarrow 1$. This procedure was repeated a total of $N = 10{,}000$ times and the relative representation r^s_{ij} for each of these realizations ($s = 1, \ldots, N$) was calculated by normalizing as explained above. The average

$$r^{\mathrm{av}}_{ij} = \sum_{s=1}^{N} r^s_{ij} / N$$

and standard deviation

$$\sigma = \sqrt{\sum_{s=1}^{N} [\left(r^{\mathrm{av}}_{ij} - r^s_{ij}\right)^2] / N}$$

were calculated as usual. The standard deviation represents an estimate of how robust the relative representation is with respect to noise, and the average is an estimate of the relative representation corrected by the addition of noise.

A P value was calculated to sort the families according to the following relevance criteria: high relative representation in the termite hindgut environment and low standard deviation. The P value for the relative representation of a family i in an environment j was calculated by adding together the number of environments for each family and generation that had a relative normalization equal to or higher than the average r^{av}_{ij} and a standard deviation equal to or smaller than the standard deviation σ_{ij} for that family and environment, divided by the total number of possible positive hits ($4NF$)

$$p_{ij} = \sum_{k=1,\ldots,F, l=1,\ldots,4, s=1,\ldots,N} Q_{kls} / 4NF$$

where $Q_{kls} = 1$ if both $r^s_{kl} \geq r^{\mathrm{av}}_{ij}$ and $\sigma_{kl} \leq \sigma_{ij}$; and zero otherwise. The comparisons were done by first binning the values of r^{av}_{ij} and σ_{ij} in intervals of size 0.01 and 0.005 respectively. In this way, a P value of, say, 0.05 for a given family means that only 5% of the families are more relevant to our analysis (higher representation in the termite environment and lower standard deviation) than the considered family.

The 1% error level is, admittedly, arbitrary. The only condition it must fulfil is that it not be large enough to distort the initial data set. A selection of different levels of noise (for example 3% or 0.1%) changes the standard deviations linearly (while the noise is kept low enough) but does not change the P value or family ordering. The pfam representation and P values were calculated in the same way as ORF families.

Hydrogenases. Analysis of abundance of hydrogenase families with different cofactor specificity was performed using the tools provided in IMG/M[32]. Genes in the termite hindgut metagenome were assigned to COG clusters and Pfam families using reverse position-specific BLAST (RPS-BLAST[57]) against the NCBI Conserved Domain Database[58] with a maximal e-value of 0.1 and a minimal percentage identity of 20%. Genes containing iron-only hydrogenase catalytic domain according to COG and Pfam classification (COG4624 and pfam02906) were clustered into families based on a minimal percentage of sequence identity of 70% for proteins within the family. Proteins within each family were aligned and consensus sequences were generated using Multalin[59], because in our experience this outperforms other tools when aligning fragmented sequences. Domain composition of each family was determined using InterProScan[60] of the consensus sequences. The catalytic domain of iron-only hydrogenases comprising proximal Fe–S clusters FS4A, FS4B and active-site H cluster[61] were aligned using ClustalW[62]; the ARB software package[48] was used for alignment editing, generation of consensus sequences for each hydrogenase family and tree calculation. Three-dimensional comparative modelling of iron-only hydrogenase families was performed using 3D-JIGSAW server[63] and was based on the structure of *Clostridium pasteurianum* iron-only hydrogenase (PDB: 1FEH[61]). Visualization of modelled structures was performed using the VMD package[64].

H_2-dependent reduction of CO_2 to acetate. For the analysis of CO_2-reductive acetogenesis summarized in Supplementary Fig. 20, Wood–Ljungdahl pathway-associated proteins were blasted against the termite metagenome using the tools provided in IMG/M. The GenBank accession numbers and species for each of the alleles employed in this search are listed here. Formate dehydrogenases and other molybdopterin enzymes: P07658 (*E. coli*) and ABC20599 (*Moorella thermoacetica*). Formyl-THF synthetase: AAP55207 (*Treponema primitia* ZAS-2). Methenyl-THF cyclohydrolase: AAP55206 (*Treponema primitia* ZAS-2). Methylene-THF dehydrogenase: YP_430368 (*Moorella thermoacetica*). Methylene-THF reductase: NP_662255 (*Chlorobium tepidum* TLS). Methyl transferase (THF to Fe–S–Co protein): AAA53548 (*Moorella thermoacetica*). Large-subunit Fe–S–Co protein: Q07340 (*Moorella thermoacetica*). Small-subunit Fe–S–Co protein: AAA23255 (*Moorella thermoacetica*). CooH-like ion translocating NiFe hydrogenase: YP_360647 (*Carboxydothermus hydrogenoformans*). RnfC-like ion translocating oxidoreductase: CAA51399 (*Rhodobacter capsulatus*). CODH/ACS CooS-subunit: P31896 (*Rhodospirillum rubrum*). CooC Ni-insertion protein: YP_430061 (*Moorella thermoacetica*). Acetyl-CoA synthase (AcsR): P27988 (*Moorella thermoacetica*). CODH/ACS-associated membrane protein with ferredoxin domain (COG3894): YP_430057 (*Moorella thermoacetica*). For the phylograms presented in Supplementary Figs 21, 22 and 24, proteins from each family were aligned using ClustalW[62]. The ARB software package[48] was used for alignment editing, generation of sequence masks and calculation of phylograms for each.

Dinitrogen fixation. For the analysis of nitrogen fixation associated genes, key proteins were blasted against the termite metagenome using the tools provided in IMG/M. The GenBank accession numbers for each of the alleles employed in this search are listed here. *Clostridium beijerinckii* genome information was used as the source of full-length *Nif*D, E, H, K and N alleles: AAS91667, AAS91669, AAF77055, AAS91668 and AAS91670, respectively. Partial *Nif*H sequences from the databases were also blasted against the metagenome. The GenBank accession numbers for each of these truncated alleles are listed here. From ref. 24: BAA28469, clone NTN2; BAA28480, clone NTN6; BAA28467, clone NTN18; BAA28470, clone NTN21; BAA28474, clone NTN30; BAA11781, clone TDY5. From ref. 65: AAK01229, *Treponema primitia* ZAS-2; AAK01231, *Treponema azotonutricium* ZAS-9; AAK01222, *Spirochaeta stenostrepta.*

Signal transduction. Distribution of genes encoding signal transduction (ST) proteins in the various metagenomic data sets and that of representative completed genomes was based on hits to COG functional categories, which include not only full-length genes but also separate domains. The search for ST COG categories was done in IMG/M using the default parameters. The abundance data were normalized using the Z-score function in the IMG

Metagenomics database, which takes into account the number of genes in the different data sets. The resulting COG categories hits were further organized into broad functional categories (for example, chemotaxis, other types of two-component system families, individual kinases, phosphatases, sensors and regulators genes or domains). The data were visualized graphically using the software Genesis[66] as a two-dimensional matrix in which the greyscale intensity is proportional to the abundance value.

28. Robertson, D. E. *et al.* Exploring nitrilase sequence space for enantioselective catalysis. *Appl. Environ. Microbiol.* **70,** 2429–2436 (2004).
29. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucleic Acids Res.* **16,** 7583–7600 (1988).
30. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
31. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
32. Markowitz, V. M. *et al.* An experimental metagenome data management and analysis system. *Bioinformatics* **22,** e359–e367 (2006).
33. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34,** D344–D348 (2006).
34. Marchler-Bauer, A., Panchenko, A. R., Ariel, N. & Bryant, S. H. Comparison of sequence and structure alignments for protein domains. *Proteins* **48,** 439–446 (2002).
35. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2,** 953–971, doi:10.1038/nprot.2007.131 (2007).
36. Wei, J. *et al.* Global proteome discovery using an online three-dimensional LC-MS/MS. *J. Proteome Res.* **4,** 801–808 (2005).
37. DeSantis, T. Z. Jr *et al.* NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* **34,** W394–W399 (2006).
38. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).
39. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21,** 456–463 (2005).
40. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18,** 502–504 (2002).
41. Schloss, P. D. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71,** 1501–1506 (2005).
42. Colwell, R. K. & Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B* **345,** 101–118 (1994).
43. de Caprariis, P., Lindemann, R. & Collins, C. A method for determining optimum sample size in species diversity studies. *J. Int. Assoc. Math. Geol.* **8,** 575–581 (1976).
44. Smith, E. P. & van Belle, G. Nonparametric estimation of species richness. *Biometrics* **40,** 119–129 (1984).
45. Krebs, C. J. *Ecological Methodology* (Harper & Row, New York, 1989).
46. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).
47. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
48. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32,** 1363–1371 (2004).
49. Austin, J. W., Szalanski, A. L. & Cabrera, B. J. Phylogenetic analysis of the subterranean termite family *Rhinotermitidae* (*Isoptera*) by using the mitochondrial cytochrome oxidase II gene. *Ann. Entomol. Soc. Am.* **97,** 548–555 (2004).
50. Ohkuma, M. *et al.* Molecular phylogeny of Asian termites (*Isoptera*) of the families *Termitidae* and *Rhinotermitidae* based on mitochondrial COII sequences. *Mol. Phylogenet. Evol.* **31,** 701–710 (2004).
51. Henrissat, B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* **280,** 309–316 (1991).
52. Davies, D. G. & Geesey, G. G. Regulation of the alginate biosynthesis gene *algC* in *Pseudomonas aeruginosa* during biofilm development in continous culture. *Appl. Environ. Microbiol.* **61,** 860–867 (1995).
53. Johnston, D. in *Handbook of Food Enzymology* (eds Whitaker, J. R., Voragen, A. G. J. & Wong, D. W. S.) 761–769 (Marcel Dekker Inc., New York, 2002).
54. Stahlberg, J., Johansson, G. & Pettersson, G. *Trichoderma reesei* has no true exo-cellulase: all intact and truncated cellulases produce new reducing end groups on cellulose. *Biochim. Biophys. Acta* **1157,** 107–113 (1993).
55. Wood, T. Preparation of crystalline, amorphous and dyed cellulase substrate. *Methods Enzymol.* **160,** 19–25 (1988).
56. von Mering, C. *et al.* STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35,** D358–D362 (2007).
57. Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32,** W327–W331 (2004).
58. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for protein classification. *Nucleic Acids Res.* **33,** D192–D196 (2005).
59. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16,** 10881–10890 (1988).
60. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).
61. Peters, J. W., Lanzilotta, W. N., Lemon, B. J. & Seefeldt, L. C. X-ray crystal structure of the Fe-only hydrogenase (CpI) from *Clostridium pasteurianum* to 1.8 angstrom resolution. *Science* **282,** 1853–1858 (1998).
62. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31,** 3497–3500 (2003).
63. Bates, P. A., Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins Struct. Funct. Genet.* **45,** 39–46 (2001).
64. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 27–28 (1996).
65. Lilburn, T. G. *et al.* Nitrogen fixation by symbiotic and free-living spirochetes. *Science* **292,** 2495–2498 (2001).
66. Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18,** 207–208 (2002).