Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen

Matthias Hess *et al.*

Science **331**, 463 (2011);
DOI: 10.1126/science.1200387

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2011 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

gous to what has been observed when predators exhibit spatial complementarity in foraging: Prey seeking to escape one predator enter habitats in which they are more vulnerable to the other predator (*26*). In addition to eliminating prey refuges, temporal partitioning may reduce the frequency of antagonistic interactions between predators, which probably explains why lizards did not affect ant abundance in our study.

When seaweed subsidies were added to the system, the greater-than-additive effect of ants and lizards on herbivory was not present. We suggest that the seaweed caused both ants and lizards to spend more time on the ground foraging for detritivores associated with seaweed deposits, reducing their combined impact on herbivory. This hypothesis is consistent with the results of an earlier study, in which seaweed was added or removed from shoreline plots on large islands (*12*). Despite higher abundances of both ants and lizards in seaweed-addition plots, *C. erectus* sustained higher levels of herbivory. An analysis of carbon-stable-isotope signatures indicated that lizard diets contained more marine-derived prey in seaweed-addition plots, and ants were observed foraging for detritivores in the seaweed deposits, suggesting that shifts in predator foraging behavior were responsible for the observed increases in herbivory (*12*). In the current experiment, the increase in ant density in pan traps may have been caused by a shift in foraging patterns, although an increase in the overall abundance of ants resulting from increased food supply may have also been contributory. Although lizard density increased in response to seaweed subsidies in mainland plots (*12*), we observed no such increase in the current study. We suggest that the absence of a numerical response in lizards on small islands was caused by (i) the inability of lizards to immigrate to the experimental sites from surrounding areas (as they could in the previous, mainland-plot experiment) and (ii) the lack of reproductive activity during much of the study period (*12*). Because emigration and reproductive lags influence the timing of predator responses to subsidy, we expect the long-term impact of pulsed seaweed subsidies on predator effects to depend on the frequency of pulses and degree of habitat isolation (*15*, *16*). Seaweed did not appear to cause a reduction in the effects of either ants or lizards by themselves (Fig. 2B), suggesting that the interactive effect of these two predators is more sensitive to subsidy than their individual effects.

Predicting the effects of environmental change on ecosystems is an important challenge. There is increasing recognition that species interactions strongly influence how environmental change affects ecosystem processes, complicating efforts to make reliable forecasts (*28*, *29*). Our results show that large seaweed-deposition events affect the structure and function of an ecological community by reconfiguring the effects of multiple predators on lower trophic levels. This suggests that predictions that are based on single-species responses or pairwise interactions may not adequately represent community responses to environmental perturbations. Experiments such as the one we report here, conducted at a spatial scale large enough to capture community-wide dynamics, are particularly relevant for conservation and management decisions in the face of ever-increasing anthropogenic disturbances.

References and Notes

1. C. Parmesan, *Annu. Rev. Ecol. Evol. Syst.* **37**, 637 (2006).
2. G. R. Walther *et al.*, *Nature* **416**, 389 (2002).
3. L. H. Yang, V. H. W. Rudolf, *Ecol. Lett.* **13**, 1 (2010).
4. G. R. Walther, *Philos. Trans. R. Soc. B* **365**, 2019 (2010).
5. B. T. Barton, A. P. Beckerman, O. J. Schmitz, *Ecology* **90**, 2346 (2009).
6. B. T. Barton, O. J. Schmitz, *Ecol. Lett.* **12**, 1317 (2009).
7. J. P. Harmon, N. A. Moran, A. R. Ives, *Science* **323**, 1347 (2009).
8. E. Post, R. O. Peterson, N. C. Stenseth, B. E. McLaren, *Nature* **401**, 905 (1999).
9. C. C. Wilmers, W. M. Getz, *PLoS Biol.* **3**, 571 (2005).
10. J. B. C. Jackson, *Proc. Natl. Acad. Sci. U.S.A.* **105** (suppl. 1), 11458 (2008).
11. H. L. Blomquist, J. H. Pyron, *Am. J. Bot.* **30**, 28 (1943).
12. D. A. Spiller *et al.*, *Ecology* **91**, 1424 (2010).
13. M. A. Bender *et al.*, *Science* **327**, 454 (2010).
14. L. B. Marczak, R. M. Thompson, J. S. Richardson, *Ecology* **88**, 140 (2007).
15. L. H. Yang *et al.*, *Ecol. Monogr.* **80**, 125 (2010).
16. R. D. Holt, *Ecology* **89**, 671 (2008).
17. G. A. Polis, W. B. Anderson, R. D. Holt, *Annu. Rev. Ecol. Syst.* **28**, 289 (1997).
18. C. V. Baxter, K. D. Fausch, M. Murakami, P. L. Chapman, *Ecology* **85**, 2656 (2004).
19. S. J. Leroux, M. Loreau, *Ecol. Lett.* **11**, 1147 (2008).
20. S. Nakano, H. Miyasaka, N. Kuhara, *Ecology* **80**, 2435 (1999).
21. J. L. Sabo, M. E. Power, *Ecology* **83**, 1860 (2002).
22. Materials and methods are available as supporting material on *Science* Online.
23. D. A. Spiller, T. W. Schoener, *Ecology* **75**, 182 (1994).
24. T. W. Schoener, D. A. Spiller, *Am. Nat.* **153**, 347 (1999).
25. J. Piovia-Scott, *J. Ecol.* **99**, 327 (2011).
26. A. Sih, G. Englund, D. Wooster, *Trends Ecol. Evol.* **13**, 350 (1998).
27. G. Takimoto, T. Iwata, M. Murakami, *Am. Nat.* **173**, 200 (2009).
28. K. B. Suttle, M. A. Thomsen, M. E. Power, *Science* **315**, 640 (2007).
29. S. E. Gilman, M. C. Urban, J. Tewksbury, G. W. Gilchrist, R. D. Holt, *Trends Ecol. Evol.* **25**, 325 (2010).

30. We thank L. H. Yang, J. J. Stachowicz, M. L. Stanton, and three anonymous reviewers for comments on the manuscript; A. N. Wright and S. S. Porter for help in the field; L. Wong for help in the lab; and the Bahamas Ministry of Agriculture and Marine Resources for permission to conduct this research. This project was supported by grants from NSF and the University of California Davis Center for Population Biology to the authors.

Supporting Online Material

www.sciencemag.org/cgi/content/full/331/6016/461/DC1
Materials and Methods
SOM Text
Figs. S1 to S3
References

10 November 2010; accepted 23 December 2010
10.1126/science.1200282

Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen

Matthias Hess,[1,2]* Alexander Sczyrba,[1,2]* Rob Egan,[1,2] Tae-Wan Kim,[3] Harshal Chokhawala,[3] Gary Schroth,[4] Shujun Luo,[4] Douglas S. Clark,[3,5] Feng Chen,[1,2] Tao Zhang,[1,2] Roderick I. Mackie,[6] Len A. Pennacchio,[1,2] Susannah G. Tringe,[1,2] Axel Visel,[1,2] Tanja Woyke,[1,2] Zhong Wang,[1,2] Edward M. Rubin[1,2]†

The paucity of enzymes that efficiently deconstruct plant polysaccharides represents a major bottleneck for industrial-scale conversion of cellulosic biomass into biofuels. Cow rumen microbes specialize in degradation of cellulosic plant material, but most members of this complex community resist cultivation. To characterize biomass-degrading genes and genomes, we sequenced and analyzed 268 gigabases of metagenomic DNA from microbes adherent to plant fiber incubated in cow rumen. From these data, we identified 27,755 putative carbohydrate-active genes and expressed 90 candidate proteins, of which 57% were enzymatically active against cellulosic substrates. We also assembled 15 uncultured microbial genomes, which were validated by complementary methods including single-cell genome sequencing. These data sets provide a substantially expanded catalog of genes and genomes participating in the deconstruction of cellulosic biomass.

Biofuels derived from lignocellulosic plant material represent an important renewable energy alternative to transportation fossil fuels (*1*, *2*). A major obstacle to industrial-scale production of fuel from lignocellulose lies in the inefficient deconstruction of plant material, owing to the recalcitrant nature of the substrate toward enzymatic breakdown and the relatively low activity of currently available hydrolytic enzymes. Although the success of protein engineering to improve the performance of existing lignocellulose-degrading enzymes has been limited (*3*), retrieving enzymes from naturally evolved biomass-degrading microbial communities offers a promising strategy for the identification of new lignocellulolytic enzymes with potentially improved activities (*4*).

Metagenomics, the direct analysis of DNA from environmental samples, represents a strategy for discovering diverse enzymes encoded in nature (*5*, *6*). Although metagenomics has been used

[1] Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA. [2] Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3] Energy Biosciences Institute, University of California, Berkeley, CA 94720, USA. [4] Illumina Inc., Hayward, CA 94545, USA. [5] Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA. [6] Department of Animal Sciences, Institute for Genomic Biology and Energy Biosciences Institute, University of Illinois, Urbana, IL 61801, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: emrubin@lbl.gov

successfully to identify enzymes with desired activities (*7*), it has relied primarily on relatively low-throughput function-based screening of environmental DNA clone libraries (*8*, *9*). Sequence-based metagenomic discovery of complete genes from environmental samples has been limited by the microbial species complexity of most environments and the consequent rarity of full-length genes in low-coverage metagenomic assemblies (*8*, *10*, *11*).

In this study, we generated 268 gigabases of metagenomic sequence data from the microbiota in cow rumen to identify genes and genomes participating in biomass deconstruction. To isolate rumen microbes associated with defined plant substrates for subsequent genomic assessment, we incubated biomass-containing nylon bags in two fistulated cows (*12*) (Fig. 1 and table S1). We isolated organisms that had become adherent to the plant fiber material during incubation in an attempt to target microbes specifically involved in biomass degradation.

We used switchgrass (*Panicum virgatum*), a promising cellulosic energy crop (*13*), as the plant substrate for our studies. To determine the cow's ability to degrade this substrate, we compared the chemical composition of the switchgrass before and after rumen incubation. Switchgrass degradation was substantial (37% dry mass reduction after 72 hours of incubation). Further analysis confirmed that the decrease in mass of the switchgrass fiber was largely due to degradation of both cellulose and hemicellulose, which together accounted for 72% of the reduction in dry mass during incubation (table S2). The remaining reduction in dry mass is likely in large part due to the degradation of pectin, protein, and other components of plant biomass (*14*). These results indicate that the cow rumen microbiota is able to degrade this fiber source and support previous observations that the rumen environment contains some of the most cellulolytic mesophilic microbes described from any habitat (*15*).

To examine whether a unique fiber-degrading microbial community was enriched on switchgrass incubated within the nylon bags, we compared the community composition of microbes adherent to rumen-incubated switchgrass to the microbial population from bulk rumen fluid. We used pyrotag sequencing of small subunit ribosomal RNA genes (*16*) to identify operational taxonomic units (OTUs) in two fistulated cows for each of the two samples. Rarefaction analysis indicated that the pyrotag sequencing depth was sufficient to capture the vast majority of OTUs in each sample and suggests that about 1000 different OTUs were present in each of these samples (fig. S1), which is consistent with previous estimates of microbial complexity in the rumen (*17*). Comparison of the OTUs identified in the switchgrass fiber-adherent community and the community present in rumen fluid revealed overlaps between cows and substrates, as well as reproducible enrichment of specific bacterial phylotypes in the switchgrass-adherent fraction (*18*).

We targeted a single sample of switchgrass-adherent rumen microbes for deep metagenomic sequencing with the goal of maximizing the likelihood of obtaining large contiguous stretches of overlapping sequence reads (contigs) containing full-length lignocellulolytic genes. We generated several sequencing libraries from this sample with paired-end read separations (equivalent to insert sizes in clone-based libraries) of 200 base pairs (bp), 300 bp, 3 kbp, and 5 kbp. Massively parallel sequencing (*19*) from all libraries yielded 1.5 billion read pairs, ranging in length from 2 × 36 bp to 2 × 125 bp and amounting to a total of 268 Gbp of sequence information. A summary of the library and sequencing technologies used is provided in table S3.

To identify candidate carbohydrate-active genes from this metagenomic sequence data set, we performed de novo assembly and predicted 2,547,270 open reading frames (ORFs). The average ORF length was 542 bp, and 55% of the ORFs were predicted to represent full-length genes. All predicted genes were screened for candidate proteins with potential enzymatic activity toward plant cell wall polysaccharides. To minimize the dependence on overall sequence similarity of candidate genes to known carbohydrate-active enzymes, we searched candidate genes for the presence of individual predicted functional domains, rather than global sequence similarity to known carbohydrate-active enzymes. We identified 27,755 candidate genes with a significant match to at least one relevant catalytic domain or carbohydrate-binding module (*18*) (table S4). The sequence domains identified in our sample were largely consistent with a prokaryotic origin of can-

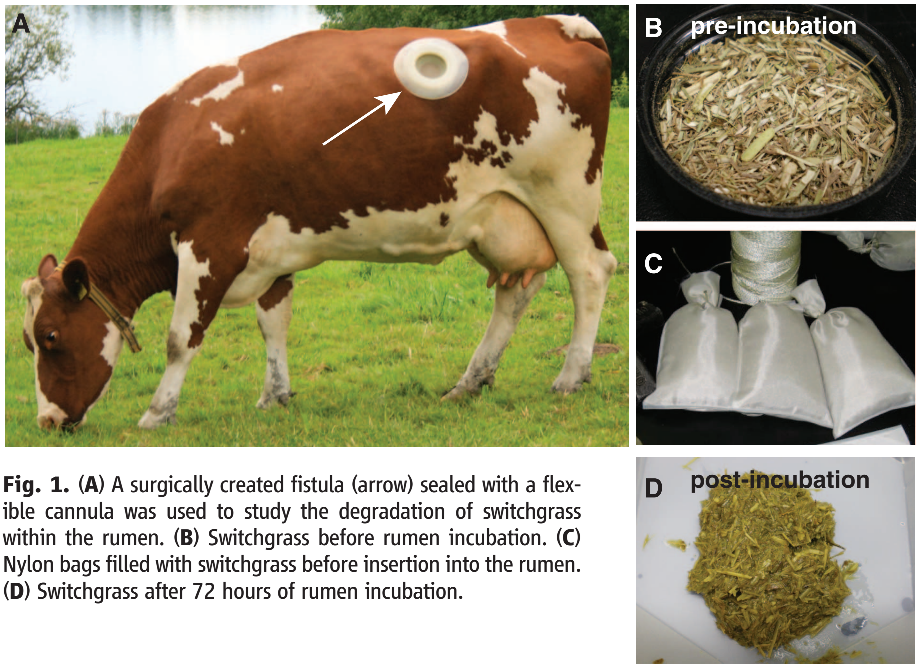


Fig. 1. (**A**) A surgically created fistula (arrow) sealed with a flexible cannula was used to study the degradation of switchgrass within the rumen. (**B**) Switchgrass before rumen incubation. (**C**) Nylon bags filled with switchgrass before insertion into the rumen. (**D**) Switchgrass after 72 hours of rumen incubation.

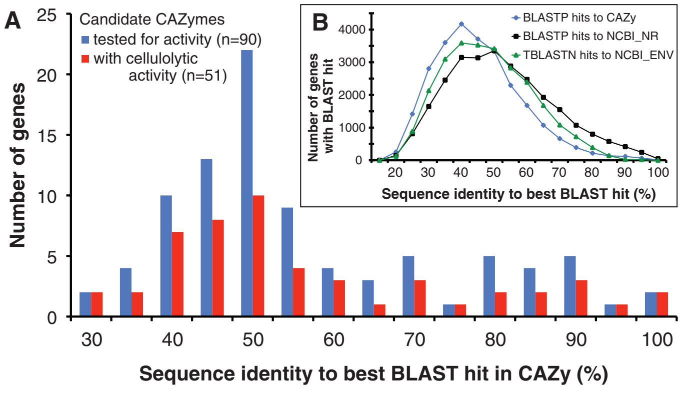


Fig. 2. (**A**) Sequence identity of 90 candidate sequences assembled from the switchgrass-associated rumen microbiome and tested for carbohydrate-degrading activity to known carbohydrate-active enzymes. Sequence identity to known enzymes is shown for tested candidates (blue) and candidates found to be active (red) toward at least one of the substrates used in the activity assays. (**B**) Similarity distribution of CAZyme candidates (*n* = 27,755) containing a catalytic domain (CD) associated with carbohydrolytic activity or a carbohydrate-binding module (CBM). Sequences were compared to the CAZy (blue, 25,947 hits), NCBI-nr (black, 26,679 hits), and NCBI-env (green, 26,030 hits) databases (best BLAST hit, E-value ≤ 1e-5); 482 genes contained both a CD and CBM, whereas 23,804 and 3469 genes contained only a CD or CBM, respectively.

didate gene sequences, with isolated examples of domains that are characteristic for eukaryotes.

Comparison of the 268 Gbp obtained by sequencing of the switchgrass-adherent microbiome to data from previously published lower-depth metagenomic studies of other plant-feeding animals (*8*, *10*, *11*) revealed that the number of candidate carbohydrate-active genes identified in the present study was larger by a factor of 5 than the combined number of candidate carbohydrate-active genes from all previous studies (table S5). The total amount of sequence analyzed in these earlier studies (combined: 0.21 Gbp) was three orders of magnitude less than the present data set and thus resulted predominantly in identification of partial genes. In contrast, genes in the present study are derived from assemblies with an average of 56-fold sequence coverage, and more than 15,000 of the candidate carbohydrate-active enzymes reported here were predicted to represent full-length genes. Rarefaction analysis indicates that even at the considerable sequencing depth of this study, only a subset of genes present in the cow rumen microbiota was assembled (figs. S4 and S5).

Although the present study focuses on the validation of a subset of carbohydrate-active enzyme families, we expect the full repertoire of genes involved in biomass deconstruction to be present in the fiber-adherent rumen metagenomic data set. To test this hypothesis, we searched our data set for cohesins and dockerins, proteins commonly involved in the formation of lignocellulolytic multi-enzyme complexes (cellulosomes) (*20*), and cellobiose phosphorylases, proteins belonging to the family of glycosyltransferases. We were able to identify 80 and 188 ORFs containing the cohesin- and dockerin-specific PFAM domains, respectively. We also identified 811 genes from the switchgrass-adherent rumen microbiome that had significant similarity to cellobiose phosphorylases deposited in NCBI-nr (BLAST search, E ≤ 1e-5). These results indicate that a wide spectrum of biomass-degrading genes can be identified through analysis of the sequence data generated in this study.

Focusing on our set of 27,755 predicted carbohydrate-active enzymes, we compared their sequences to entries in the Carbohydrate Active Enzyme (CAZy) database, which contains both experimentally verified and inferred carbohydrate-active enzymes (*21*). In the CAZy database, 1075, 1199, and 251 entries are annotated as β-(1,4)

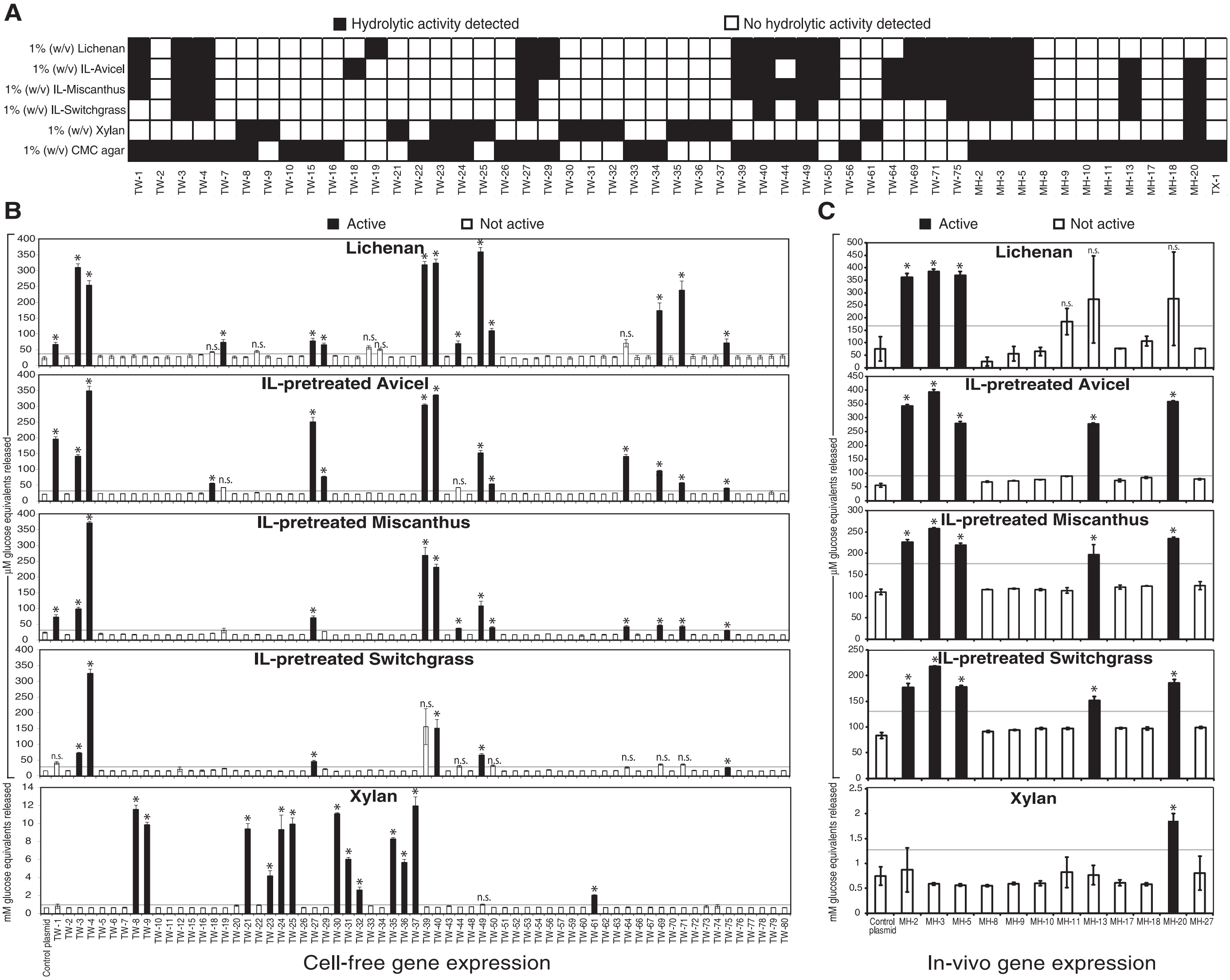


Fig. 3. Carbohydrolytic potential of candidate carbohydrate-active enzymes on glycosidic substrates of different complexity. (**A**) Summary of carbohydrolytic activities of 90 tested candidates on ten substrates. Candidates that were not active on any of the tested substrates and the four substrates that were recalcitrant to all tested candidates are not shown. (**B** and **C**) Cellulolytic activities of candidate carbohydrate-active enzymes on five substrates. Carbohydrate-active gene candidates were expressed in a cell-free system (B) or using *Escherichia coli* as expression host (C). Samples were only considered "active" (■) if the measured mean glucose equivalent quantitatively exceeded the activity of negative controls plus one standard deviation by at least 50% (indicated by shaded horizontal line in each panel) and was significantly higher (*$P < 0.05$, Student's *t* test) than the negative controls. Samples not meeting both criteria were considered as "not active" (□; n.s. = not significant). All measurements were performed in duplicate. IL, ionic liquid.

endoglucanases, β-glucosidases, and cellobiohydrolases, respectively (63%, 86%, and 87% of these entries lack an Enzyme Commission (EC) number, indicating that their assigned activity has not been verified biochemically). In our rumen-derived data set, we identified 1086, 1477, and 153 sequences whose most significant matches (BLAST search, E ≤ 1e-5) were to a β-(1,4) endoglucanase, β-glucosidase, or cellobiohydrolase, respectively, within the CAZy database. Only 1% of these 2716 sequences were highly similar (>95% sequence identity) to any CAZy database entry, indicating that nearly all of these enzymes had not been previously deposited in CAZy. The overall lower efficiency at which new candidate cellobiohydrolases were identified may be due to their underrepresentation in the reference database, but even in this category we observed a 56% increase of candidate sequences with <95% sequence identity to sequences previously deposited in the CAZy database.

Conventional sequence homology–based enzyme discovery introduces a bias toward the identification of candidates similar to known enzymes, rather than new enzymes with low sequence identity and potentially divergent physicochemical properties. To assess our ability to discover carbohydrate-active enzymes with limited overall sequence identity to known proteins, we compared the amino acid sequences of the 27,755 putative carbohydrate-active genes identified in our metagenomic data set to all genes deposited in the NCBI nonredundant (NCBI-nr) and environmental database (NCBI-env) and to all 85,740 carbohydrate-active enzymes deposited in CAZy. Only 12% of the 27,755 carbohydrate-active genes assembled from the rumen metagenome were more than 75% identical to genes deposited in NCBI-nr, whereas 43% of the genes had less than 50% identity to any known protein (Fig. 2B). Twenty-four percent of our candidate carbohydrate-active genes were most similar to sequences annotated as "hypothetical protein" or "predicted protein" in NCBI-nr. Moreover, only 5% of our carbohydrate-active enzyme sequences are more than 75% identical to sequences deposited in the NCBI-env database (Fig. 2B and fig. S6), demonstrating that these enzymes also have not been observed in previous metagenome projects. These results reveal the abundance and diversity of putative carbohydrate-active enzymes in the fiber-adherent rumen microbiome.

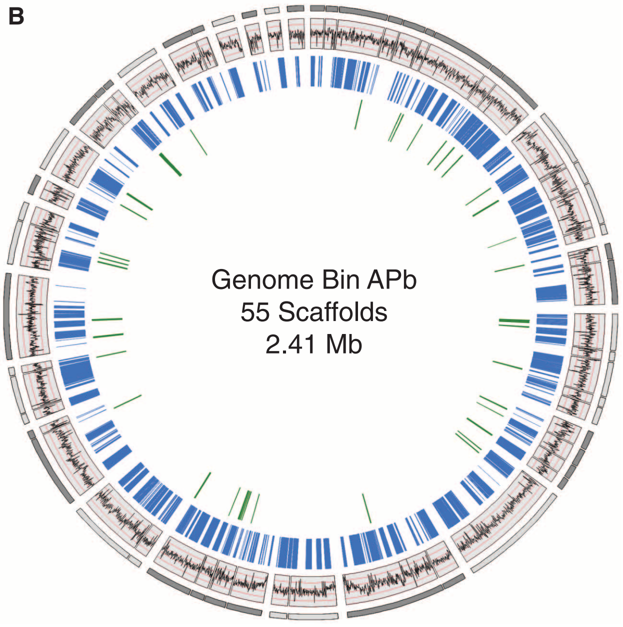
To examine the validity of gene assemblies from short-read sequence data, we experimentally investigated a random subset ($N = 233$) of the putative rumen carbohydrate-active genes. We designed gene-specific primer pairs for each of these predicted carbohydrate-active genes and attempted to amplify the corresponding targets from the same DNA used to generate the metagenomic data. Using a single set of amplification conditions, we obtained polymerase chain reaction (PCR) products with the predicted size for 158 of 233 (68%) candidate genes. A randomly selected subset of the PCR products was further validated by sequencing, which confirmed in nearly all cases the computationally predicted sequence of the assembled candidate genes (>95% sequence identity, 28 of 29 putative genes). These results suggest that a substantial proportion of the genes predicted on the basis of short-read assemblies extracted from the metagenomic data represent authentic genes present in rumen microbes.

To evaluate the biochemical activity of the putative carbohydrate-active genes identified by metagenomic sequencing of the switchgrass-associated microbiome, we chose 90 candidate genes predicted to contain a glycoside hydrolase family 3, 5, 8, 9, 10, 26, or 48 domain or a carbohydrate-binding module. The selected candidate genes were expressed using two complementary expression systems, and the obtained proteins were subjected to biochemical activity assays. The genes selected for expression ranged from 29 to 96% amino acid sequence identity to known carbohydrate-active proteins, with an average of less than 55% identity (Fig. 2A). We tested all 90 proteins for enzymatic activity on a panel of 10 different substrates. This panel included eight model substrates—carboxymethyl cellulose (CMC), *p*-nitrophenyl β-glucoside, gum guar, lichenan, laminarin, mannopentose, Avicel, and xylan—along with two potential biofuel feedstocks, miscanthus and switchgrass (*22*), to provide an initial understanding of the substrate specificity for each of the tested candidates. The biofuel crop substrates and Avicel were subjected to ionic liquid pretreatment before conducting the activity assays. In total, 51 of 90 (57%) tested proteins showed enzymatic activity against at least one of the substrates, suggesting that the candidate genes predicted by our metagenomic strategy are highly enriched in enzymes with relevant activities (Fig. 3 and table S6). There was no evidence that proteins with high sequence identity to known enzymes were more likely to be active than proteins with low sequence similarity ($P = 0.66$, Kolmogorov-Smirnov test; Fig. 2A). Inactivity of the remaining carbohydrate-active candidates in these assays could be due to a number of reasons, including false-positive prediction of carbohydrate-active enzyme domains, minimal expression and/or misfolding of candidate proteins, or suboptimal reaction conditions. The overall high validation rate observed in these assays suggests that the number and sequence diversity of known genes encoding hydrolytic enzymes from these and possibly other enzyme

Fig. 4. (**A**) Draft genomes assembled from switchgrass-adherent rumen microbes. Completeness was estimated as fraction of the number of identified and the number of expected core genes within the phylogenetic order. For more information on the assembled genomes, see table S10. (**B**) Circular representation of the draft genome (genome bin *APb*) validated by single amplified genome analysis. From outside toward the center: outermost circle, scaffolds within the draft genome (random order, low-quality regions removed); circle 2, Illumina read coverage in metagenomic data (each horizontal red line indicates 25-fold coverage); circle 3 (blue tick marks), regions of the draft genome simultaneously covered by 454 reads derived from a single amplified genome; innermost circle (green tick marks), location of glycoside hydrolase genes on draft genome.

A

Genome Bin	Genome Size (Mb)	Phylogenetic Order	Estimated Completeness
AFa	2.87	Spirochaetales	92.98%
AMa	2.21	Spirochaetales	91.23%
Ala	2.53	Clostridiales	90.10%
AGa	3.08	Bacteroidales	89.77%
AN	2.02	Clostridiales	78.50%
AJ	2.24	Bacteroidales	75.96%
AC2a	2.07	Bacteroidales	75.96%
AWa	2.02	Clostridiales	75.77%
AH	2.52	Bacteroidales	75.45%
AQ	1.91	Bacteroidales	71.36%
AS1a	1.75	Clostridiales	70.99%
APb	2.41	Clostridiales	64.85%
BOa	1.67	Clostridiales	64.16%
ADa	2.99	Myxococcales	62.13%
ATa	1.87	Clostridiales	60.41%



families was substantially increased through this metagenomic data set.

A considerable fraction (20%) of the tested carbohydrate-active enzyme candidates showed activity toward biofuel crops pretreated with ionic liquids, one of the most promising initial steps in the deconstruction of biomass (*23*). Because ionic liquids can inhibit enzymatic biomass degradation (*24*), the retention of enzymatic activity in their presence makes these proteins promising candidates for more detailed physicochemical analyses. In addition, the tested set of target candidates also included two enzymes (MH-9 and MH-10) that showed activity on CMC agar plates and contained only a carbohydrate-binding module, but no known catalytic domain specific for carbohydrate-active enzymes (table S6). It is possible that these two enzymes contain catalytic domains that share little similarity with the catalytic domains of currently known carbohydrate-active enzymes.

To enable genomic studies of these microbes, we developed a strategy for producing draft genomes from deep metagenomic data. An initial assembly of the 268 Gbp of metagenomic sequence resulted in 179,092 scaffolds, of which the 65 largest ranged in size from 0.5 to 1.5 Mbp (tables S8 and S9). Only 47 (0.03%) of the assembled scaffolds showed high levels of similarity (≥90% identity over ≥1000 bp) to previously sequenced genomes available in GenBank. Most of these alignable scaffolds were small (median length: 1626 bp) and the aligned regions typically covered nearly the entire length of the scaffold (median: 91% of scaffold length). These results suggest that the vast majority of the assembled scaffolds represent segments of hitherto uncharacterized microbial genomes. We further validated these assemblies via two independent indicators of scaffold integrity: (i) level and uniformity of read depth in subregions, and (ii) mate-pair support. We identified 26,042 scaffolds greater than 10 kbp that satisfied these criteria for scaffold integrity (*18*), totaling 568 Mbp (N50: 24 kbp; longest scaffold: 541 kbp). To generate draft genomes, we binned the validated scaffolds by means of two complementary properties expected to be present in scaffolds derived from the same genome: (i) tetranucleotide frequencies (TNFs) and (ii) read coverage. TNF signatures are generally an effective approach for distinguishing sequences derived from different genomes but can be similar for closely related species (*25*, *26*). In contrast, read coverage is directly correlated to the relative abundance of each organism in the sample and can thus be used to distinguish scaffolds that are likely derived from different closely related organisms. In total, 446 genome bins with consistent TNF and read coverage were formed. To estimate the completeness of the largest potential microbial draft genomes identified through this approach, we first determined the most likely phylogenetic order from which each of these bins was derived. For each of these orders, we used all available sequenced reference genomes (table S11) to identify a minimal set of core genes that are present in all members of this order (*27*, *28*). Comparison of each draft genome to the pangenome of the respective phylogenetic order demonstrated that between 60% and 93% of the core genes were included in the 15 draft genomes found to be most complete by this measure, similar to the fraction found in each reference genome used for comparison (Fig. 4A and tables S10 and S12). These observations suggest that near-complete draft genomes were successfully assembled. To address the possibility that the completeness of individual draft genomes was overestimated as a result of binning of scaffolds derived from multiple organisms, we further validated their authenticity by copy number analysis of genes that were present only in single copy in all reference genomes of the respective phylogenetic order (*18*) (tables S10 and S12).

To test experimentally the validity and completeness of draft genomes derived from metagenomic scaffold bins, we obtained genome sequence data from individual uncultured microbial cells isolated directly from the same complex rumen community. Single cells loosely adherent to switchgrass were isolated using fluorescence-activated cell sorting (*29*) followed by whole genome amplification (*30*). Screening of 16*S* sequences suggested that one of the single cells analyzed was related to the fibrolytic *Butyrivibrio fibrisolvens* and matched bin *APb*, one of the largest bins assembled from metagenomic data (Fig. 4A and table S10). From this particular single cell, we generated 65,272 reads (22.5 Mbp after appropriate filtering) of which 55% mapped to genome bin *APb*. The remaining mappable single-cell reads matched either unbinned scaffolds or assembly regions with poor scaffold integrity. Each of the 55 scaffolds in bin *APb* was supported by substantial numbers of mapped single cell–derived reads, suggesting that all scaffolds in bin *APb* represent segments of the genome of the same single organism (Fig. 4B). These results directly support the assumption that individual genome bins derived from our assembly represent authentic draft genomes and suggest that substantial proportions of the respective genomes are covered by these bins.

Discovery of full-length genes with defined functions from complex microbial communities has previously been severely limited by the low throughput of the required cellular and molecular manipulations (*8*, *11*, *31*). Our study demonstrates the potential of deep sequencing of a complex community to accurately reveal genes of interest at a massive scale, and to generate draft genomes of uncultured novel organisms involved in biomass deconstruction. Although this work focused on the identification and validation of new carbohydrate-active enzymes, these data sets provide an extensive resource for the discovery of a multitude of other classes of enzymes known to exist in the rumen, and the general approach presented here will be applicable to other environmental microbial communities.

References and Notes

1. H. W. Blanch *et al.*, *ACS Chem. Biol.* **3**, 17 (2008).
2. K. Sanderson, *Nature* **444**, 673 (2006).
3. F. Wen, N. U. Nair, H. Zhao, *Curr. Opin. Biotechnol.* **20**, 412 (2009).
4. E. M. Rubin, *Nature* **454**, 841 (2008).
5. M. Ferrer *et al.*, *Environ. Microbiol.* **7**, 1996 (2005).
6. F. Wang, F. Li, G. Chen, W. Liu, *Microbiol. Res.* **164**, 650 (2009).
7. L. L. Li, S. R. McCorkle, S. Monchy, S. Taghavi, D. van der Lelie, *Biotechnol. Biofuels* **2**, 10 (2009).
8. P. B. Pope *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14793 (2010).
9. T. Uchiyama, K. Miyazaki, *Curr. Opin. Biotechnol.* **20**, 616 (2009).
10. J. M. Brulc *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1948 (2009).
11. F. Warnecke *et al.*, *Nature* **450**, 560 (2007).
12. J. H. Meyer, R. I. Mackie, *Appl. Environ. Microbiol.* **51**, 622 (1986).
13. D. J. Parrish, J. H. Fike, *Methods Mol. Biol.* **581**, 27 (2009).
14. W. J. Kelly *et al.*, *PLoS ONE* **5**, e11942 (2010).
15. P. J. Weimer, *J. Dairy Sci.* **79**, 1496 (1996).
16. Z. Liu, C. Lozupone, M. Hamady, F. D. Bushman, R. Knight, *Nucleic Acids Res.* **35**, e120 (2007).
17. D. O. Krause *et al.*, *FEMS Microbiol. Rev.* **27**, 663 (2003).
18. See supporting material on *Science* Online.
19. D. R. Bentley *et al.*, *Nature* **456**, 53 (2008).
20. M. T. Rincon *et al.*, *PLoS ONE* **5**, e12476 (2010).
21. B. L. Cantarel *et al.*, *Nucleic Acids Res.* **37** (database issue), D233 (2009).
22. T. Demura, Z. H. Ye, *Curr. Opin. Plant Biol.* **13**, 299 (2010).
23. H. Zhao, G. A. Baker, J. V. Cowins, *Biotechnol. Prog.* **26**, 127 (2010).
24. H. Zhao *et al.*, *J. Biotechnol.* **139**, 47 (2009).
25. H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, F. O. Glöckner, *Environ. Microbiol.* **6**, 938 (2004).
26. T. Woyke *et al.*, *Nature* **443**, 950 (2006).
27. P. Lapierre, J. P. Gogarten, *Trends Genet.* **25**, 107 (2009).
28. D. Medini, C. Donati, H. Tettelin, V. Masignani, R. Rappuoli, *Curr. Opin. Genet. Dev.* **15**, 589 (2005).
29. R. Stepanauskas, M. E. Sieracki, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9052 (2007).
30. T. Woyke *et al.*, *PLoS ONE* **4**, e5299 (2009).
31. H. García Martín *et al.*, *Nat. Biotechnol.* **24**, 1263 (2006).
32. We thank J. Bristow, P. Hugenholtz, F. Warnecke, and K. Mavrommatis for critical discussions and reading the manuscript. We acknowledge technical support by the JGI production team, L. M. Sczyrba, M. Harmon-Smith, J. Froula, J. Martin, C. Wright, A. Lipzen, J. Zhao, S. Malfatti and Stefan Bauer. We thank P. D'Haeseleer for sequences extracted from the CAZy database, Jonas Løvaas Gjerstad for the picture of the fistulated cow, T. Shinkei, T. Yannarell, J. Kim and staff at the Dairy Farm, Department of Animal Sciences for assistance with the maintenance of the fistulated cows, nylon bag experiments and lab procedures carried out at the University of Illinois. The work conducted by the U.S. Department of Energy Joint Genome Institute was supported in part by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH112 and U.S. Department of Energy under contract DE-AC02-05CH11231 (cow rumen metagenomics data analysis and informatics). Supported by a research grant from the Energy Biosciences Institute at the University of California, Berkeley (M.H.). Data are available at the NCBI Short Read Archive under accession number SRA023560 and GenBank accession numbers HQ706005-HQ706094. Complete data can also be accessed through the Web site of the DOE Joint Genome Institute (www.jgi.doe.gov).

Supporting Online Material

www.sciencemag.org/cgi/content/full/331/6016/463/DC1
Materials and Methods
SOM Text
Tables S1 to S13
Figs. S1 to S8
References

12 November 2010; accepted 29 December 2010
10.1126/science.1200387