J. Gen. Appl. Microbiol., **64**, 108–116 (2018)
doi 10.2323/jgam.2017.08.004

jGAM

Full Paper

Metagenomic insights into lignocellulose-degrading genes through Illumina-based *de novo* sequencing of the microbiome in Vietnamese native goats' rumen

(Received December 5, 2016; Accepted August 21, 2017; J-STAGE Advance publication date: March 12, 2018)

Thi Huyen Do,[1,3] Ngoc Giang Le,[1,2] Trong Khoa Dao,[1,3] Thi Mai Phuong Nguyen,[1] Tung Lam Le,[1] Han Ly Luu,[1] Khanh Hoang Viet Nguyen,[3,4] Van Lam Nguyen,[1,3] Lan Anh Le,[1] Thu Nguyet Phung,[1] Nico M. van Straalen,[2] Dick Roelofs,[2] and Nam Hai Truong[1,3,*]

[1] *Institute of Biotechnology, Vietnam Academy of Science and Technology, 18-Hoang Quoc Viet, Cau Giay, Ha Noi, Vietnam*
[2] *Department of Ecological Science, Vrije Universiteit Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands*
[3] *Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18-Hoang Quoc Viet, Cau Giay, Ha Noi, Vietnam*
[4] *Institute of New Technology/Academy of Military Science and Technology, 17 Hoang Sam Street, Nghia Do, Cau Giay, Hanoi, Vietnam*

The scarcity of enzymes having an optimal activity in lignocellulose deconstruction is an obstacle for industrial-scale conversion of cellulosic biomass into biofuels. With the aim of mining novel lignocellulolytic enzymes, a ~9 Gb metagenome of bacteria in Vietnamese native goats' rumen was sequenced by Illumina platform. From the data, 821 ORFs encoding carbohydrate esterases (CEs) and polysaccharide lyases (PLs) serving for lignocellulose pre-treatment, 816 ORFs encoding 11 glycoside hydrolase families (GHs) of cellulases, and 2252 ORFs encoding 22 GHs of hemicellulases, were mined. The carbohydrate binding module (CBM) was also abundant with 763 ORFs, of which 480 ORFs are located with lignocellulolytic enzymes. The enzyme modularity analysis showed that CBMs are usually present in endoglucanase, endo 1,3-beta-D-glucosidase, and endoxylanase, whereas fibronectin 3-like module (FN3) mainly represents in GH3 and immunoglobulin-like domain (Ig) was located in GH9 only. Every domain located in each ORF was analyzed in detail to contribute enzymes' modularity which is valuable for modelling, to study the structure, and for recombinant production. With the aim of confirming the annotated results, a mined ORF encoding CBM63 was highly expressed in *E. coli* in soluble form. The purified recombinant CBM63 exhibited no cellulase activity, but enhanced a commercial cellulase activity in the destruction of a paper filter.

Key Words: carbohydrate binding model; cellulase; hemicellulase; Illumina *de novo* sequencing; metagenome; pretreatment; Vietnamese native goat

Abbreviations: CEs, carbohydrate esterases; CBM, carbohydrate binding module; FN3, fibronectin 3-like module; GHs, glycoside hydrolase families; Ig, immunoglobulin-like domain; ORFs, open reading frames; PLs, polysaccharide lyases; KEGG, Kyoto Encyclopedia of Genes and Genomes; eggNOG, evolutionary genealogy of genes; Non-supervised Orthologous Groups; COG, Cluster of Orthologous Groups; CAZy, Carbohydrate-Active enZYmes; GO, Gene Ontology; ARDB, antibiotic resistance genes database

Introduction

Lignocellulose waste comprising agro-industrial biomass is inexpensive, renewable, abundant, and provides a unique natural resource for enhancing bio-economy (Anwar et al., 2014) to substitute the fossil-based economy. Overcoming the limitations of fossil-based economy, bio-

*Corresponding author: Nam Hai Truong, Institute of Biotechnology, Vietnam Academy of Science and Technology, 18-Hoang Quoc Viet, Cau Giay, Ha Noi, Vietnam.
Tel: +84-43-791-7980 E-mail: tnhai@ibt.ac.vn

None of the authors of this manuscript has any financial or personal relationship with other people or organizations that could inappropriately influence their work.

based economy has the advantage to i) be environmentally, economically and socially sustainable; (ii) decrease the dependence on fossil fuel; (iii) reduce atmospheric greenhouse gas emission, which is responsible for causing climate change; and (iv) stimulate regional and rural development (de Jong et al., 2012). Lignocellulose can be converted into sugar molecules by microbial enzymes and the released sugars can be fermented into various high value products including bio-fuels, materials for food, bulk chemicals such as bioplastics, and value-added fine chemicals for pharmaceuticals and human health (Asgher et al., 2013; Iqbal et al., 2013; Irshad et al., 2012; Millati et al., 2011). Therefore, lignocellulose biomass has recently gained increasing research interest and special importance (Asgher et al., 2013; Baumann et al., 2016; Ofori-Boateng and Lee, 2013).

The conversion of lignocellulose into higher-value products requires a multi-step process including (i) pre-treatment (e.g. mechanical, chemical, or biological), (ii) saccharification by enzymes, and (iii) fermentation into end products (Arumugam and Mahalingam, 2015). A major obstacle to lignocellulose conversion in industry lies in the inefficient deconstruction of plant material owing to the retention of the natural lignocellulose structure. Also, currently available enzymes which can hydrolyze lignocellulose show a low and ineffective activity (Hess et al., 2011; Sebastian et al., 2013). In nature, individual enzymes interact synergistically, or are comprised of multi-modules (modularity), to degrade lignocellulose effectively. In modularity, besides the catalytic core, these enzymes also possess non-catalytic functionally-important domains, including carbohydrate-binding modules (CBMs), fibronectin 3-like modules (FN3s), dockerins, immunoglobulin-like domains (Ig), or functionally unknown "X" domains (Sweeney and Xu, 2012). These domains are important for solubility, optimal activity (Ding et al., 2008; Wilson, 2008), stability and even thermo-stability of the catalytic activity (Araki et al., 2006; Jia et al., 2016). In *Clostridia*, these enzyme modules are organized in so-called cellulosomes through cohensin-dockerin complexes (Dou et al., 2015). Apparently, organisation and interaction of these microbial enzymes for the hydrolysis of lignocellulose are essential in the industrial development of lignocellulose breakdown, which is an important source for the green energy sector (Kumar et al., 2016; Yang et al., 2014). Many recent studies have identified numerous potentially enzymatic pathways for biomass conversion, but less is known about the efficacy of catalytic activity of the enzyme modularity in biomass transformation and digestion (Kumar et al., 2016). Thus, the discovery of novel enzyme modularity for lignocellulose saccharification is required.

Traditionally, functional microbial screening is applied to isolate genes involved in lignocellulose breakdown. More recently, metagenomics can identify candidate genes from environmental samples circumventing the need for culturing. This is important, since more than 99% of microorganisms from environmental samples are uncultivable and their functional significance is overlooked. Thus, next-generation sequencing of whole metagenomic DNA from environmental samples with a high lignocellulose breakdown capacity is very powerful for the discovery of genes relevant in this process (Kumar et al., 2016; Sebastian et al., 2013). The digestive tract of termites (Do et al., 2014; Kumar et al., 2016; Sebastian et al., 2013), and Korean goat rumen (Lim et al., 2013) represent rapid and efficient lignocellulose degradation environments, which make it more likely to discover enzymes that play an essential role in this process. Much emphasis has been given to investigating enzymes from microbiota that can hydrolyse cellulose, and hemicellulose substrates. However, much less information is available on the collocation of important domains (FN3, CBM and Ig) forming modules with catalytic domains to eventually create an efficient system for optimal lignocellulose degradation. Lim et al. (2013) reported nine CBM domains, dockerin-1, and FN3 domains, and these domains were collocated within cellulase and glycosyl hydrolase (GH) families, but lacked all information on genes for many hemicellulases and genes for lignocellulose pretreatment. In addition, most identified cellulases lacked a co-localized with CBM and/or FN3 domain (Lim et al., 2013).

Here, we report on the analysis of a large dataset generated by Illumina-based *de novo* sequencing of bacterial metagenomic DNA extracted from the rumen of native goats living in the natural high mountain at Ninh Binh and Thanh Hoa, Vietnam. These animals consume different plant materials with a high content of lignocellulose. Therefore, we hypothesize that the microbial digestive system of this animal has adapted to degrade substantial amounts of lignocellulose efficiently. A previous study used only one database to analyze goat rumen to identify potentially relevant enzymes (Lim et al., 2013). In this study, we have subjected all open reading frames (ORFs) to six available functional annotation tools. This integrated approach increased the number of identified cellulases and hemicellulases, and enzymes related to lignocellulose pretreatment. We have also analyzed the presence of collocated FN3, CBM, and Ig domains, thereby elucidating the potential of an enzyme to participate in modularity. This information may become necessary for the recombinant production of optimal enzyme cocktails.

Materials and Methods

Ethics statement. The animal experimental protocol of this research was reviewed, discussed, and approved by the institutional ethics committee of the Institute of Genome Research Institutional Review Board (IGR IRB), Vietnam (Approval number: No. 03/QD-NCHG).

Sampling and extraction of bacterial metagenomic DNA. The goat lines used in this study were a Vietnamese native breed (Co) and a hybrid (Bach Thao) generated by Beetal and Jamnapari long time ago. Adult Co animals, weigh approximately 30 to 35 kg (Fig. S1A) and live on natural hay in high rocky mountains at private goat farms at Ninh Binh and Thanh Hoa provinces in Vietnam. The domestic goat breed Co has a small body with brown or black hair, a large head, small short ears, and short horns. The breed Bach Thao is diverse in morphology and size (Fig. S1B). Three Co animals and two Bach Thao animals were sampled in Ninh Binh province (GPS coordinates

20.269002 105.893267), while two Co animals and three Bach Thao animals were sampled in Thanh Hoa province (GPS coordinates 19.897450 105.795899). The diet of both goat lines consists of a variety of grasses, leaves of trees in the mountains, and also crop residues at night.

In total, ten selected goats were slaughtered at a local slaughter house. Rumen fluid from each goat was filtered through four layers of cheesecloth, and the remains was suspended in 2 liters of PBS buffer (137 mM NaCl, 2.0 mM KH_2PO_4, 10 mM Na_2HPO_4, and 2.7 mM KCl, pH 7.4). It was filtered through a new set of four layers of cheesecloth. The resulting fluids were centrifuged at 700 rpm (approximately 150–200 g) for 10 min to separate protozoa and plant debris from bacteria. This step was repeated twice. The bacteria in the supernatant were pelleted by centrifugation (4,500 *g* for 5 min), washed twice with PBS buffer, and resuspended in 500 ml of PBS buffer.

Genomic DNA was isolated from the bacteria-enriched fluid and purified using a PSP Spin Stool DNA Plus Kit (Stratec, Germany) according to the manufacturer's protocol. The extracted DNA was checked by agarose gel electrophoresis, quantified and quality-checked by NanoDrop ND-2000C (Implen, US) before storage. Equal amounts of DNA from the 10 goat rumens were mixed for sequencing. The mixed metagenomic DNA showed only slight degradation, and was concentrated to 132 μg/ml (OD260/280 value of 1.92). Total 10 μg of the DNA was sent to BGI-Hong Kong Co. Ltd. for sequencing.

Metagenome sequencing and assembly. The paired-end library was prepared as described elsewhere (Do et al., 2014). The metagenomic DNA was sequenced using next generation ultra high throughput sequencing system Illumina HiSeq2500 (Illumina, San Diego, CA, USA). The raw sequence data was analyzed using a standard bioinformatics approach as follows. Adaptor sequences and reads containing >10% "N" bases, and reads containing >50% low quality base scores (Q < 20), were removed from the raw data. The reads were then assembled by SOAPdenovo2 (Luo et al., 2012) with different k-mer sizes in parallel, and Rabbit tool (You et al., 2013) was used to extend the length of SOAPdenovo-derived contigs. Reads were then mapped back to the final contigs for each assembly in order to choose the most optimal k-mer size and to select the best assembly with regard to N50 and coverage. Contigs with a length of >200 bps were kept for open reading frame (ORF) prediction using MetaGeneMark (Zhu et al., 2010). The predicted genes were further clustered using CD-hit (Li and Godzik, 2006). The genes having a sequence identity ≥95% and alignment coverage ≥95% were merged and kept for functional annotation.

Functional annotation. All the predicted ORFs were blasted against public databases: (i) Swiss-Prot; (ii) Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2008); (iii) Non-supervised Orthologous Groups (eggNOG); (iv) Cluster of Orthologous Groups (COG) (Powell et al., 2011); (v) Carbohydrate-Active enZYmes (CAZy) database (Cantarel et al., 2009); and (vi) Gene Ontology (GO) (Ashburner et al., 2000). A flow chart of the bioinformatics pipeline for analysis of the bacterial metagenomic DNA extracted from Vietnamese native goats' rumen is represented in Fig. S2. Top hits, those with an E-value lower than 10^{-5} and a sequence coverage >50%, and the highest sequence similarity, were used for further analysis. Lignocellulolytic enzyme families (Pfam, protein families) were predicted by performing Interproscan (https://www.ebi.ac.uk/interpro/).

For taxonomic classification, homology of the mapped ORFs was queried to previously characterized ORFs in the non-redundant (NR) database in NCBI. In this way, organism diversity was obtained in the goat rumens at the phylum level. For the annotation of species, the best matching ORFs, whose E value was lower than e^{-5}, were preserved in the classified group for further analysis. The ORFs in the classified group were subjected to MEGAN (version 4.6) (Huson et al., 2007) for assignment into NCBI taxonomy using the lowest common ancestor (LCA) algorithm.

This project was deposited in the DNA Databank of Japan (DDBJ) with the accession ID PSUB006562.

CBM63 expression, purification and activity analysis. For assessment of the functional annotated results, the ORF 57823 encoding CBM63 was chosen for *E. coli* gene expression and activity analysis.

This gene (858 bps) contains a 5′ terminal sequence of 78 bps encoding a signal peptide and another sequence spanning the remaining 777 bps codes for a mature CBM63. The gene encoding mature CBM63 was synthesized by Genescript (USA) and cloned into pET22b(+) (Novagen) at *Nco*I and *Xho*I restriction sites. The obtained plasmid was introduced into *E. coli* BL21 (DE3) (Novagen). For protein expression, a single-colony transformant was inoculated into 10 ml Luria-Bertani broth (supplemented with 100 μg/ml ampicillin; LBA), and grown overnight at 37°C in a rotary shaker (200 rpm). The overnight culture (0.2 ml) was then transferred to 20 ml of fresh LBA and cultivated at 37°C, 200 rpm until the optical density (OD_{600}) reached 0.6–0.8. Subsequently, the cells were induced for CBM63 expression by adding 0.5 mM IPTG and continuously grown for 4 hours at 25°C. The cells were harvested by centrifugation at 6,000 rpm for 10 min at 4°C, and suspended in water to a density of OD_{600} = 10. The cells were disrupted by sonication on an ice bath (10 pulses, 30 s each at 100 W with 20 s intermission). The soluble fraction was separated from the pellet by centrifugation at 13,000 rpm for 10 min at 4°C. The expressed proteins in soluble and insoluble fractions were analyzed by SDS-PAGE. The recombinant CBM63 was purified by Immobilized Metal Affinity Chromatography (IMAC) with a 5 mL Ni-charged Sepharose Fast Flow column (HisTrap; GE Healthcare). Before loading the sample, the column was equilibrated with 10 column volumes (CV) of buffer (20 mM KH_2PO_4, 0.5 M NaCl, pH 7.4) containing 50 mM imidazole. After applying the soluble fraction to the column, it was washed with 5 CV of the same buffer containing 100 mM imidazole, and eluted by 10 CV of the buffer containing 500 mM imidazole. The protein concentration in the purified fractions was measured by NanoDrop ND-2000C (Implen, US) and was checked by electrophoresis SDS-PAGE and then desalted using a PD10 desalting column (GE Healthcare).

The purified CBM63 was used to check the activity with carboxymethyl cellulose (CMC, Sigma) and filter paper as substrates. Whatman No. 1 filter paper was cut into very small pieces by scissors. The total reaction volume was 0.5 ml containing: 10 mg of the filter paper (or 0.1 mg CMC); 0.05 ml of 0.05 M Na-citrate buffer, pH 6; and 0.3 mg purified CBM63 protein with, or without, 0.025 U of cellulase (Sigma). The reaction was performed at 50°C for exactly 90 min and then stopped immediately by adding 0.5 ml of dinitro salicylic (DNS) reagent. All the tubes were boiled for 5 min and the absorbance at 540 nm was measured. Each measurement was made in triplicate. The activity of the protein was calculated as the amount of reducing sugar (corresponding to mM glucose in this study) released (Miller, 1959).

Results and Discussion

Sequencing analysis

Illumina sequencing of the metagenomic DNA yielded 89,964,640 reads. Of these, 84,625,346 reads (94.07%) were useful reads used for assembly to 172,918 contigs larger than 200 bp by a SOAPdenovo assembly tool using a k-mer size of 51. From the contigs, 164,644 ORFs were predicted (Table S1). The inventory of ORFs length distribution is shown in Fig. S3.

Similarity searches using BLAST against the non-redundant protein sequence database showed that 122,304 ORFs (74.3%) retrieved a Blast hit. The nrBLAST output was subjected to MEGAN (version 4.6) (Huson et al., 2007) for taxonomic assignment. Among 39,579 ORFs affiliated in taxonomic classification, most of the genes (99.8%) originated from bacteria. Only nine ORFs belonged to Eukaryota, two ORFs originated from viruses and 67 ORFs were classified to Archaea (Fig. 1). This confirms the enrichment of bacterial DNA during the metagenomic DNA extraction of goat rumen samples.

Among the bacterial genes, phylum Bacteroidetes was the most represented, accounting for 63.6%, followed by Firmicutes (22.6%), and Proteobacteria (7.5%) (Fig. 1). Also, these phyla are most abundantly represented in the microbial eco-system in Japanese goat rumens (Denman et al., 2015). Earlier studies showed that the dominance of Bacteroidetes is correlated to the presence of cellulolytic glycoside hydrolases (GH), which play an important role in lignocellulose degradation (Güllert et al., 2016; Han et al., 2015). The dominance of Bacteroidetes may reflect high lignocellulolytic degradation activity in the goat rumen.

For functional annotation, the ORFs were blasted against diverse databases. In total, 141,521 ORFs were annotated. Typical eukaryotic COG categories, such as RNA processing and modification, chromatin structure were almost not represented in our data set (Fig. S4). This result supports again the observation that our metagenomic DNA extraction was highly enriched for bacterial DNA.

The number of ORFs matching to each of the COG, eggNOG, KEGG, GO, CAZy, and Swiss-Prot, databases were 37,007 (Fig. S4), 134,843; 56,751; 86,693; 7,898 and 33,471 ORFs, respectively. However, in this study we are specifically interested in gene functions related to carbohydrate metabolism. As such, 3,642 genes could be annotated to the COG category carbohydrate transport and metabolism, while 17,984 ORFs received this annotation with eggNOG. Moreover, a subset of this gene set (11,999 ORFs) could be identified to be involved in the carbohydrate metabolism category in KEGG. As expected, almost all genes annotated by CAZy databases (7898 ORFs) were valuable for mining carbohydrate degrading enzymes.

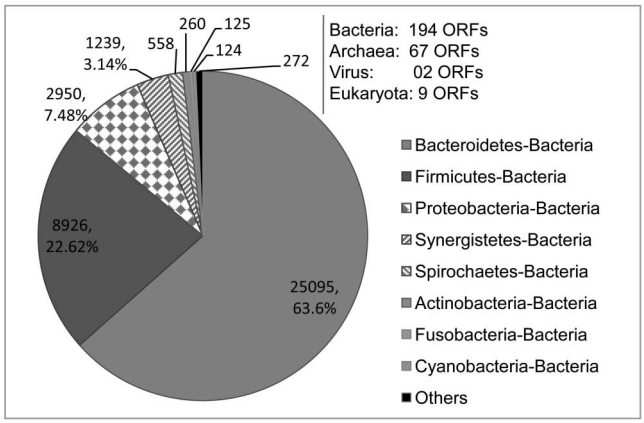

Fig. 1. Analysis of the goat rumen microbial community structure at the phylum level.

The numbers of ORFs affiliated in each phylum and its percentage are indicated however percentage is not indicated for less than 3.14%.

Functional annotation showed an abundance of ORFs encoding putative enzymes/proteins for lignocellulose degradation

The CAZy annotation exhibited mainly four kinds of catalytic domain related to carbohydrate degradations that comprised GHs (4,715 ORFs), glycosyl transferases (GTs: 1,956 ORFs), polysaccharide lyases (PLs: 229 ORFs), and carbohydrate esterases (CEs: 969 ORFs). The 4,715 ORFs classified in GHs categories were divided into 65 GH families (Table 1). According to CAZy, within these families, 11 GHs belonged to cellulases (Table S2), 22 GHs belonged to hemicellulases (Table S3), and 32 GHs contain other activity domains. Unfortunately, no enzyme responsible for lignin-degradation, such as Mn-peroxidase or laccase, was found. An earlier study showed that the lignin degradation process in ruminants is usually limited in rumen, due to the anaerobic conditions (Susmel and Stefanon, 1993). Microflora in animal rumen is constituted by facultative anaerobic bacteria ($1–10 \times 10^9$ per ml) protozoa and fungi, however fungi are found to play a predominant role in lignin degradation (Kasuya et al., 2007; Susmel and Stefanon, 1993), while bacteria and protozoa are responsible for the efficient degradation of cellulose and hemicellulose (Moreira et al., 2013; Susmel and Stefanon, 1993). In addition, lignin was revealed to have a positive function in the rumen to help maintain the reservoir of buffering exchangeable cations for feed digestion (Moreira et al., 2013).

Carbohydrate esterase families (CE and PL) are known to enhance lignocellulose pretreatment. The ORFs encoding these families were found in abundance in our data with a total of 821 ORFs. Most CEs were related to pectin

Table 1. Summary of CAZy annotation of the genes from bacterial metagenomic DNA extracted from Vietnamese native goats' rumen.

Name	ORFs	Name	ORFs	Name	ORFs	Name	ORFs	Name	ORFs	Name	ORFs
CBM:	763	CE:	969	GH:	4715	GH2	372	GH5	192	GT:	1956
CBM0	11	CE1	257	GH0	30	GH20	40	GH51	138	GT0	28
CBM2	13	CE10	163	GH1	16	GH23	105	GH53	79	GT1	7
CBM3	3	CE11	47	GH10	116	GH24	37	GH57	45	GT10	4
CBM4	11	CE12	104	GH103	3	GH25	109	GH63	1	GT11	40
CBM6	122	CE13	3	GH105	112	GH26	98	GH64	1	GT19	45
CBM9	2	CE14	2	GH106	46	GH27	19	GH66	2	GT2	933
CBM13	31	CE15	35	GH108	6	GH28	210	GH67	58	GT23	1
CBM20	66	CE2	33	GH109	4	GH29	67	GH73	62	GT26	19
CBM22	2	CE3	1	GH11	2	GH3	400	GH74	1	GT28	60
CBM25	2	CE4	66	GH112	1	GH30	16	GH77	115	GT3	35
CBM32	62	CE6	105	GH113	1	GH31	152	GH78	65	GT30	52
CBM34	6	CE7	68	GH115	121	GH32	61	GH8	48	GT32	20
CBM35	26	CE8	75	GH119	1	GH33	42	GH84	3	GT35	74
CBM37	56	CE9	10	GH120	4	GH35	75	GH88	7	GT4	397
CBM38	2	PL:	229	GH125	4	GH36	52	GH89	17	GT41	2
CBM41	2	PL1	108	GH127	62	GH38	1	GH9	46	GT5	63
CBM48	127	PL10	36	GH13	326	GH4	2	GH91	1	GT51	111
CBM50	205	PL11	76	GH130	44	GH42	1	GH92	37	GT8	6
CBM57	9	PL9	9	GH16	35	GH43	641	GH94	50	GT83	6
CBM61	4			GH18	14	GH44	2	GH95	115	GT9	53
CBM63	1			GH19	3	GH48	1	GH97	178		

esterase, only CE6 was suggested to be reductase and carboxylesterase. We identified 61 ORFs that contain both hemicellulase (GH10), as well as esterase (CE1), domains. In addition, another 19 ORFs encoded bifunctional domains. Within this group, 18 ORFs encoded both hemicellulase GH26 and esterase CE7, while 1 ORF encoded a protein with hemicellulase GH43 and esterase CE6 domains (Table S4). The enzyme with a bifunctional domain may be useful for application because, at the same time, two activities can be synergistically exhibited simultaneously for improving the substrate degradation (Neddersen and Elleuche, 2015). The ORFs divided into PL groups in this study mostly have catalytic domains for pectin degradation (Table S4).

For cellulose degradation, the functional annotation has assigned 816 ORFs encoding cellulases, which were categorized in 11 GHs (Table S2). While, according to CAZy, GH16, GH5, GH8, GH9 were related to endoglucanase, GH3 was beta-glucosidase, and GH16 was suggested to be glucan endo-1,3-beta-D-glucosidase and licheninase. For hemicellulose degradation, after integration of COG, KEGG and GO annotated results, a total of 2252 ORFs were predicted to encode hemicellulases, including endo-1,4-beta-xylanase, beta-xylosidase, and 20 kinds of branching enzymes (Table S3).

Besides the catalytic core, many of lignocellulases possess non-catalytic, but functionally important, domains for their activity. These domains include CBM, FN3, dockerins, Ig, and so-called unknown "X" domains. CBM has an affinity to an individual or bundled polysaccharide chains, as well as to single carbohydrate molecules. Thus, it anchors or directs host enzymes to targeted carbohydrate substrates (Guillén et al., 2010). In some cases, CBM exerts the ability to disrupt crystalline cellulose microfibrils to assist cellulase reactions (Ding et al., 2008; Wilson, 2008). In this study, 763 ORFs harbouring domains of 21 types of CBM, including a CBM63 (which may possess expansin activity to disrupt the crystal structure of lignocellulose), were mined (Table S5). In this, 15 types of CBMs (480 ORFs) were collocalized with cellulase (9 ORFs), and hemicellulase domains (241 ORFs) (Tables S2 and S3). Interestingly, all CBMs collocated with endoglucanase catalytic domain and endo 1,3-beta-D-glucosidase catalytic domain, but never co-localized with beta-glucosidase domain (which accounted for ~50% predicted cellulases). This suggests that, during cellulose degradation, endoglucanase first opens up the cellulose structure and subsequently digests cellulose into cellobiose and other small polysaccharides. Apparently, this enzyme needs CBM for more affinity to the substrate to function more optimally. Overall, CBM domains presented in 10% ORFs encoded hemicellulases and 1% ORFs encoded cellulases. In a previous study, Dai et al. (2012) also described 10% of the plant cell wall-targeting GH proteins carrying a CBM. CBM4 and CBM22 have the capacity to bind to xylan and beta-1,3/beta-1,4-glucans, while CBM22 has a thermo-stabilizing effect for catalytic domains (Araki et al., 2006). Interestingly, CBM4 domain was identified in CE1, and CBM22 was collocated with CE3. In the group of hemicellulases having CBM domains, endo-1,4-beta-xylanase accounted for 30.6% (23 ORFs). Thus, the presence of CBM domain is clearly associated with enzymes participating in the first step of lignocellulose degradation for enhancing the enzyme affinity to more effectively decompose substrate.

The fibronectin-3-like module is known to loosen up the cellulose surface, and may separate cellulose chains and expose additional sites of cellulose for hydrolysis by the covalently-attached catalytic domain (Kataeva et al., 2002). In our study, 214 ORFs with FN3 domains were observed to be collocated with GH5 (1 ORF), and GH3 domains (213 ORFs for beta-glucosidase) (Table S2). This is in agreement with the finding in a previous study that beta-glucosidase did not harbour CBM but contained an

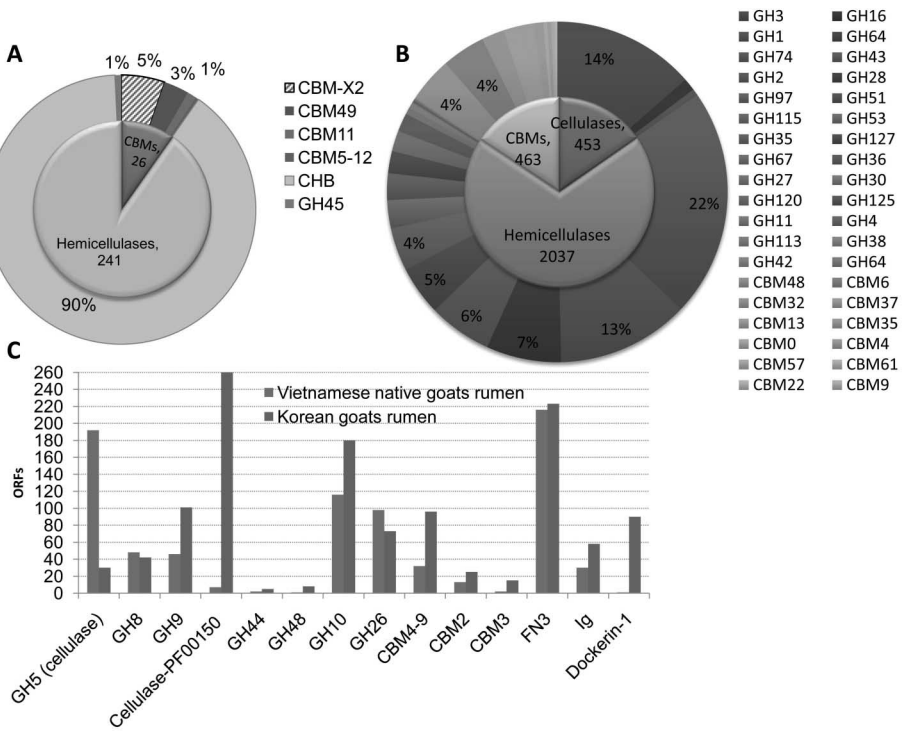

Fig. 2. Comparison of modules located in lignocellulolytic enzymes in the metagenomic data of Vietnamese native goats rumen and Korean goats rumen.

Carbohydrate-binding modules (CBM), glycoside hydrolases (GHs) for cellulases and hemicellulases, and other unfunctional domains located in lignocellulolytic enzymes annotated by CAZy and Interpro, that were observed only in the metagenomic data of Korean goats' rumen (A) (Lim et al., 2013) or exhibited only in Vietnamese native goats' rumen (B) and found in both metagenomic data of Vietnamese and Korean goat rumen bacteria (C).

FN3 domain (Sweeney and Xu, 2012). Another previous study showed that bacterial FN3 sequences were identified only in extracellular matrix proteins (Kataeva et al., 2002). This suggests that many bacterial beta-glucosidases are secreted into the goat's rumen, playing an important role in the transformation of cellulose to glucose as a nutrient for the goat, rather than providing a carbon source for bacterial consumption.

In this study, we also identified Ig domains (30 ORFs) responsible for stabilization and enhanced thermo-stability of collocated catalytic domains, accompanied by only GH9 catalytic domains. This association is confirmed by another study, where Ig plays a vital role in activating GH9 enzymes (Kataeva et al., 2004).

Comparison of metagenomic data from Vietnamese and Korean goats' rumen in the emphasis of the ORFs for putative lignocellulases

We compared ORFs data encoding cellulases to the data published by Lim et al. (2013), and found that the endoglucanase GH8 was present in both datasets in comparable abundance, while GH44 and GH48 were also present in both datasets, although at a low abundance (Fig. 2) (Lim et al., 2013). This result may reflect the presence of a well-defined group of cellulase GHs that have evolved as a specific adaptation to the specific digestive circumstances in goat rumen. However, these two studies also differ considerably. For instance, GH9, GH44, and GH10 represent endoglucanases that were identified in both Korean and Vietnam goats' rumen, although at a lower abundance (~1/2 times) in our sample. The same pattern is observed in the case of cellulase PF00150, where a 37 times greater abundance was identified in the Korean goat rumen as compared with the Vietnamese goat rumen. In contrast, GH5, which is responsible for endoglucanase, showed a 6.4 times greater abundance in the Vietnam data (Fig. 2). Some GHs were only observed in the Korean goat rumen data, but were absent in the Vietnamese goat rumen data. Whereas, many GHs for cellulase activity were observed in the Vietnamese goat rumen sequences, but were absent in Korean goat rumen data. For instance, endoglucanase GH45 and cellobiose hydrolase (CBH) were identified only in the Korean goat rumen, with a total of 241 genes. Meanwhile, we observed 453 ORFs assigned to the families GH1, GH16, GH3, GH64, and GH74, and comprising endoglucanase, beta-glucosidase, 6-phospho-beta-glucosidase, beta-glucosidase-related glycosidase, cellobiose phosphorylase, glucan endo-1,3-beta-D-glucosidase, and licheninase activities, only in the Vietnamese goat rumen data (Table S2, Fig. 2). Overall, the total genes annotated by the same databases (CAZy) for cellulases were 749 genes in our study and 687 genes in the Korean goat rumen data (Fig. 2). The size of the Korean goat rumen data is 2.4 fold greater than the Vietnamese goat rumen metagenomic data. These results indicate that bacterial cellulase genes in the rumen of the Vietnamese native goats are more abundant than those in the Korean goats. In addition, in our study, some ORFs could

not be annotated into the GH family, but were still predicted to have an activity linked to beta-glucosidase, cellulase M/endoglucanase, and endoglucanase. The total number of ORFs assigned from all databases for cellulases were 816 ORFs (Table S2). The difference in cellulase genes may lie in the bacterial sources. Several studies have provided evidence that the rumen microbiome can be influenced and shaped by the host genotype (An et al., 2005; Hess et al., 2011; Kittelmann and Janssen, 2011; Nelson et al., 2003; Sundset et al., 2007), diet preference (Han et al., 2015; Tajima et al., 2001; Zhu et al., 2014), as well as the habitat (Sundset et al., 2007). In the case of the host genotype, the Korean goats used for mining lignocellulase genes represent the Saanen hybrid line, and in this study we used genotypes derived from Co and Bach Thao hybrid lines. In general, these genotypes are omnivorous animals, feeding mainly on natural plants, leaves, agricultural waste such as straw, cornstalks, and sugarcane tops. However, we chose goats living in a mountainous area and feeding particularly on various plant and agriculture waste.

Although the overall abundance of cellulase genes is comparable to the rumen data of Vietnamese and Korean goat rumen data, the distribution of specific GH enzymes differs considerably between the two studies. This supports the notion that effective hydrolyzation of cellulases in any lignocelluose-degrading ecosystem is highly diverse and cannot be linked to a specific group of catalytic domains represented by a defined set of enzymes (Hu et al., 2013; Liu et al., 2013; Tiwari et al., 2013).

According to the CAZy annotation, 22 GHs having hemicellulase activities were found (Table S3). However, only GH10 and GH26 were observed in both metagenomic data from Korean and Vietnamese goat rumen. Overall, the absolute number of genes belonging to GH10 and GH26 in Korean goat rumen data (~256 ORFs) was slightly higher than in Vietnamese goat rumen data (214 ORFs). In contrast, the other 20 GHs, which accounted for 2037 ORFs, were observed in our data but not described in the Korean dataset (Fig. 2). This suggests that bacteria in Vietnamese goat rumen adapted specifically to the digestion of diverse lignocellulose materials in the tree and dry crop residues, which may be harsher to digest when compared with digesting lignocellulose present in young leaves.

Of the 2252 ORFs predicted to have hemicellulase activities, 20 kinds of branching enzymes were identified (Table S3). Remarkably, all the branching enzymes were absent in the Korean goat rumen dataset (Lim et al., 2013). The high abundance of hemicellulases in our metagenomic dataset may be explained by the specific diet requirements of Vietnamese native goat breeds.

The CEs and PLs were not represented at all in the bacterial metagenomic data from Korean goats rumen (Lim et al., 2013), indicating that the present dataset from Vietnam goat rumen is more diverse in the number and function of genes. The number of CEs and PLs genes affiliated to Bacteroidetes were approximately 16 times higher than that affiliated to Firmicutes. Detailed results will be published in the future.

The four most abundant CBMs (CBM6, CBM50, CBM48, CBM32) of the 21 CBM types in our data were also identified to be the four most abundant CBMs in cow rumen (Hess et al., 2011). When comparing our data with data from Korean goat rumen (Lim et al., 2013), CBM2, CBM3, and CBM4-9 were identified in both datasets, but their abundance was much lower among the Vietnamese sequences. Other CBM domains (CBM5-12, CBMX-2, CBM11, CBM19) were completely lacking in our data. In contrast, 12 CBMs among 463 ORFs were annotated only in the Vietnamese dataset. In total, 510 ORFs were annotated in our data, which was threefold higher compared with the CBM-containing genes in the Korean dataset (162 genes) (Fig. 2). Bacterial expansin is usually found in strains belonging to *Bacillus subtilis* (Kerff et al., 2008) and *Hahella chejuensis* (Lee et al., 2010), which are involved in disrupting the crystal structure of lignocellulose, enabling other cellulases to further depolymerize the liberated polysaccharides. After an extensive search in our data, we found only one gene for expansin that was annotated to be CBM63 according to CAZy. Finally, it is worth mentioning that expansin was not identified and described in Korean goat rumen (Lim et al., 2013), again indicating the more diverse and rich content of the Vietnamese goat rumen microbiome.

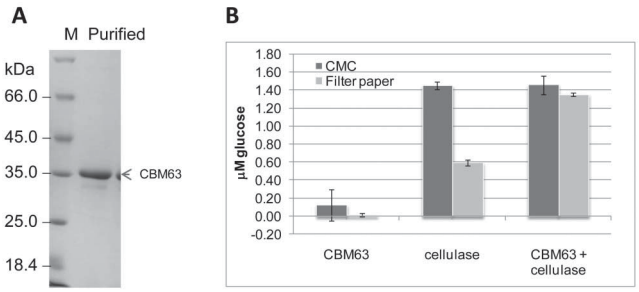

Fig. 3. Expression of CBM63 protein in *E. coli*.

SDS-PAGE analysis of purified CBM63 from recombinant *E. coli* extract (A), and assessment of CBM63 ability to enhance cellulase activity by DNS method (B). M: Standard proteins (Fermentas).

In agreement with the previous study in goat rumen, dockerin type I was only annotated in GH9, supporting previous observations (Borne et al., 2013; Hirano et al., 2015; Lim et al., 2013). Dockerin type I only exists in cellulosome modularity (Borne et al., 2013; Hirano et al., 2015). The low abundance of dockerin type I in this sample indicates that a cellulosome structure is not established in the Vietnamese goat rumen microbiome. This is supported by the fact that we also did not find any cohensin, dockerin type II in this data, which is essential for cellulosome assembly.

In Korean goat rumen metagenomic data, no clear correlation was found between FN3 and a specific catalytic domain (Lim et al., 2013).

Expression of CBM63 for preliminary confirmation of annotated results

With regard to the confirmation of functional annotated results of the genes from metagenomic data, a nucleotide sequence of 777 bp encoding for mature CBM63 was expressed in *E. coli*. By MEGAN analysis, the CBM63 was assigned to be originated from *Ruminococcus flavefaciens*.

In the amino acid sequence, CBM63 was the most closely identical with expansin of *Clostridium* sp Marseille-P2415 NCBI (WP_077613372.1, 45%) and *Bacillus atrophaeus* NCBI (WP_061669738.1, 43%). CBM63 also possesses a conserved catalytic domain of endoglucanase at the C terminus. In *E. coli*, a substantial part of the expressed CBM63 (30 kDa) was soluble and highly accumulating in *E. coli* cells. The recombinant protein was successfully purified by His-tag affinity column (Fig. 3). The purified and desalted CBM63 did not exhibit endoglucanase activity to digest CMC but was capable of significantly enhancing commercial cellulase activity to convert filter paper (a typical crystal lignocellulose) into reducing sugars as detected by DNS reagent.

With the purpose of confirming the functional annotated results of the genes from metagenomic data, Hess et al. (2011) mined 27,755 putative carbohydrate-active genes from cow rumen's metagenomic data and expressed 90 candidate proteins which had an amino acid sequence identity to known carbohydrate-active proteins ranging from 26% up to 96%. They discovered that 57% recombinant proteins exhibited enzymatic activity. There was no link between enzymatic activity with the degree of amino acid sequence identity (Hess et al., 2011). In agreement with this study, we also expressed seven other cellulose-, hemicellulose-, pectin-active genes in *E. coli*, of which five showed enzymatic activities and the remaining enzymes were expressed at too low a level (data will be published elsewhere). This indicates that the majority of mined genes possess actual activity.

In conclusion, we were able to annotate a wide diversity of hemicellulase genes that are associated with CBMs in our samples. We also observed CBMs located in cellulases and enzymes for lignocellulose pretreatment, but to a much lesser extent. The FN3 domain was in high abundance and showed a clear association with GH3, while the Ig domain was more linked to GH9. This resource will be highly useful, when recombinant enzyme assays are needed to be applied as cocktail enzymes to accomplish a more optimal industrial degradation of lignocellulose.

Acknowledgments

We would like to acknowledge Dr. S. V. N. Vijayendra (Food Microbiology Dept., Central Food Technological Research Institute, Mysore-570020, India) for proofreading and correcting the English of this manuscript. The study was carried out with the financial support of the Project "Metagenome of some potential mini-ecologies for mining novel genes encoding effective lignocellulolytic enzymes" code DTDLCN.15/14, managed by the Ministry of Science and Technology, Vietnam, in collaboration with the Department of Ecological Science, Vrije Universiteit Amsterdam, The Netherlands, supported by the BE-BASIC consortium project numbers F07.003.05 and F07.003.07. We thank the National Key Laboratory of Gene Technology, Institute of Biotechnology, VAST, Vietnam, for the use of their facilities. We are also grateful to the Editor of JGAM for valuable comments to improve the manuscript.

Supplementary Materials

Supplementary figures and tables are available in our J-STAGE site (http://www.jstage.jst.go.jp/browse/jgam).

References

An, D., Dong, X., and Dong, Z. (2005) Prokaryote diversity in the rumen of yak (*Bos grunniens*) and Jinnan cattle (*Bos taurus*) estimated by 16S rDNA homology analyses. *Anaerobe*, **11**, 207–215.

Anwar, Z., Gulfraz, M., and Irshad, M. (2014) Agro-industrial lignocellulosic biomass a key to unlock the future bio-energy: A brief review. *J. Radiat. Res. Appl. Sci.*, **7**, 163–173.

Araki, R., Karita, S., Tanaka, A., Kimura, T., and Sakka, K. (2006) Effect of family 22 carbohydrate-binding module on the thermostability of Xyn10B catalytic module from *Clostridium stercorarium*. *Biosci. Biotechnol. Biochem.*, **70**, 3039–3041.

Arumugam, N. N. and Mahalingam, P. U. (2015) Lignocellulose plant biomass; an emerging alternative fuel resource. *ResearchGate*, **XLIX**, 291–295.

Asgher, M., Ahmad, Z., and Iqbal, H. M. N. (2013) Alkali and enzymatic delignification of sugarcane bagasse to expose cellulose polymers for saccharification and bio-ethanol production. *Ind. Crops Prod.*, **44**, 488–495.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H. et al. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.

Baumann, I., Westermann, P., Baumann, I., and Westermann, P. (2016) Microbial production of short chain fatty acids from lignocellulosic biomass: current processes and market. *BioMed Res. Int.*, **2016**, doi:10.1155/2016/8469357.

Borne, R., Bayer, E. A., Pagès, S., Perret, S., and Fierobe, H. P. (2013) Unraveling enzyme discrimination during cellulosome assembly independent of cohesin-dockerin affinity. *FEBS J.*, **280**, 5764–5779.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.*, **37**, D233–D238.

Dai, X., Zhu, Y., Luo, Y., Song, L., Liu, D. et al. (2012) Metagenomic insights into the fibrolytic microbiome in Yak rumen. *PLoS ONE*, **7**, e40430.

de Jong, E., Higson, A., Walsh, P., and Wellisch, M. (2012) Biobased Chemicals - value added products from biorefineries. In IEA Bioenergy - Task 42 Biorefinery, - NNFCC.

Denman, S. E., Martinez Fernandez, G., Shinkai, T., Mitsumori, M., and McSweeney, C. S. (2015) Metagenomic analysis of the rumen microbial community following inhibition of methane formation by a halogenated methane analog. *Front. Microbiol.*, **6**, 1087.

Ding, S. Y., Xu, Q., Crowley, M., Zeng, Y., Nimlos, M. et al. (2008) A biophysical perspective on the cellulosome: new opportunities for biomass conversion. *Curr. Opin. Biotechnol.*, **19**, 218–227.

Do, T. H., Nguyen, T. T., Nguyen, T. N., Le, Q. G., Nguyen, C. et al. (2014) Mining biomass-degrading genes through Illumina-based *de novo* sequencing and metagenomic analysis of free-living bacteria in the gut of the lower termite *Coptotermes gestroi* harvested in Vietnam. *J. Biosci. Bioeng.*, **118**, 665–671.

Dou, T. Y., Luan, H. W., Ge, G. B., Dong, M. M., Zou, H. F. et al. (2015) Functional and structural properties of a novel cellulosome-like multienzyme complex: efficient glycoside hydrolysis of water-insoluble 7-xylosyl-10-deacetylpaclitaxel. *Sci. Rep.*, **5**, 13768.

Guillén, D., Sánchez, S., and Rodríguez-Sanoja, R. (2010) Carbohydrate-binding domains: multiplicity of biological roles. *Appl. Microbiol. Biotechnol.*, **85**, 1241–1249.

Güllert, S., Fischer, M. A., Turaev, D., Noebauer, B., Ilmberger, N. et al. (2016) Deep metagenome and metatranscriptome analyses of microbial communities affiliated with an industrial biogas fermenter, a cow rumen, and elephant feces reveal major differences in carbohydrate hydrolysis strategies. *Biotechnol. Biofuels*, **9**, 121.

Han, X., Yang, Y., Yan, H., Wang, X., Qu, L. et al. (2015) Rumen bacterial diversity of 80 to 110-day-old goats using 16S rRNA sequencing. *PLoS ONE*, **10**, e0117811.

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H. et al. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, **331**, 463–467.

Hirano, K., Nihei, S., Hasegawa, H., Haruki, M., and Hirano, N. (2015) Stoichiometric assembly of the cellulosome generates maximum synergy for the degradation of crystalline cellulose, as revealed by In vitro reconstitution of the *Clostridium thermocellum* cellulosome. *Appl. Environ. Microbiol.*, **81**, 4756–4766.

Hu, J., Arantes, V., Pribowo, A., and Saddler, J. N. (2013) The synergistic action of accessory enzymes enhances the hydrolytic potential of a

"cellulase mixture" but is highly substrate specific. *Biotechnol. Biofuels*, **6**, 112.

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

Iqbal, H. M. N., Kyazze, G., and Keshavarz, T. (2013) Advances in the valorization of lignocellulosic materials by biotechnology: an overview. *BioResources*, **8**, 3157–3176.

Irshad, M. N., Anwar, Z., But, H. I., Afroz, A., Ikram, N. et al. (2012) The industrial applicability of purified cellulase complex indigenously produced by *Trichoderma viride* through solid-state bioprocessing of agro-industrial and municipal paper wastes. *BioResources*, **8**, 145–157.

Jia, X., Qiao, W., Tian, W., Peng, X., Mi, S. et al. (2016) Biochemical characterization of extra- and intracellular endoxylanse from thermophilic bacterium *Caldicellulosiruptor kronotskyensis*. *Sci. Rep.*, **6**, 21672.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M. et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Kasuya, N., Wada, I., Shimada, M., Kawai, H., and Itabashi, H. (2007) Effect of presence of rumen protozoa on degradation of cell wall constituents in gastrointestinal tract of cattle. *Anim. Sci. J.*, **78**, 275–280.

Kataeva, I. A., Seidel, R. D., Shah, A., West, L. T., Li, X. L. et al. (2002) The fibronectin type 3-like repeat from the *Clostridium thermocellum* cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. *Appl. Environ. Microbiol.*, **68**, 4292–4300.

Kataeva, I. A., Uversky, V. N., Brewer, J. M., Schubot, F., Rose, J. P. et al. (2004) Interactions between immunoglobulin-like and catalytic modules in *Clostridium thermocellum* cellulosomal cellobiohydrolase CbhA. *Protein Eng. Des. Sel. PEDS*, **17**, 759–769.

Kerff, F., Amoroso, A., Herman, R., Sauvage, E., Petrella, S. et al. (2008) Crystal structure and activity of *Bacillus subtilis* YoaJ (EXLX1), a bacterial expansin that promotes root colonization. *Proc. Natl. Acad. Sci. USA*, **105**, 16876–16881.

Kittelmann, S. and Janssen, P. H. (2011) Characterization of rumen ciliate community composition in domestic sheep, deer, and cattle, feeding on varying diets, by means of PCR-DGGE and clone libraries. *FEMS Microbiol. Ecol.*, **75**, 468–481.

Kumar, M., Varma, A., and Kumar, V. (2016) Ecogenomics based microbial enzyme for biofuel industry. *Sci. Int.*, **4**, 1–11.

Lee, H. J., Lee, S., Ko, H. J., Kim, K. H., and Choi, I. G. (2010) An expansin-like protein from *Hahella* chejuensis binds cellulose and enhances cellulase activity. *Mol. Cells*, **29**, 379–385.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lim, S., Seo, J., Choi, H., Yoon, D., Nam, J. et al. (2013) Metagenome analysis of protein domain collocation within cellulase genes of goat rumen microbes. *Asian-Australas. J. Anim. Sci.*, **26**, 1144–1151.

Liu, M., Gu, J., Xie, W., and Yu, H. (2013) Directed co-evolution of an endoglucanase and a β-glucosidase in *Escherichia coli* by a novel high-throughput screening method. *Chem. Commun. Camb. Engl.*, **49**, 7219–7221.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W. et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, **1**, 18.

Millati, R. I., Syamsiah, S., Niklasson, C., Cahyanto, M. N., Lundquist, K. et al. (2011) Biological pretreatment of lignocelluloses with white-rot fungi and its applications: A review. *ResearchGate*, **6**, 5224–5259.

Miller, G. L. (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal. Chem.*, **31**, 426–428.

Moreira, L. M., de Leonel, F. P., Vieira, R. A. M., and Pereira, J. C. (2013) A new approach about the digestion of fibers by ruminants. *Rev. Bras. Saúde E Produção Anim.*, **14**, 382–395.

Neddersen, M. and Elleuche, S. (2015) Fast and reliable production, purification and characterization of heat-stable, bifunctional enzyme chimeras. *AMB Express*, **5**, 33.

Nelson, K. E., Zinder, S. H., Hance, I., Burr, P., Odongo, D. et al. (2003) Phylogenetic analysis of the microbial populations in the wild herbivore gastrointestinal tract: insights into an unexplored niche. *Environ. Microbiol.*, **5**, 1212–1220.

Ofori-Boateng, C. and Lee, K. T. (2013) Sustainable utilization of oil palm wastes for bioactive phytochemicals for the benefit of the oil palm and nutraceutical industries. *Phytochem. Rev.*, **12**, 173–190.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M. et al. (2011) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.

Sebastian, R., Kim, J. Y., Kim, T. H., and Lee, K. T. (2013) Metagenomics: a promising approach to assess enzymes biocatalyst for biofuel production. *Asian J. Biotechnol.*, **5**, 33–50.

Sundset, M. A., Praesteng, K. E., Cann, I. K. O., Mathiesen, S. D., and Mackie, R. I. (2007) Novel rumen bacterial diversity in two geographically separated sub-species of reindeer. *Microb. Ecol.*, **54**, 424–438.

Susmel, P. and Stefanon, B. (1993) Aspects of lignin degradation by rumen microorganisms. *J. Biotechnol.*, **30**, 141–148.

Sweeney, M. D. and Xu, F. (2012) Biomass converting enzymes as industrial biocatalysts for fuels and chemicals: recent developments. *Catalysts*, **2**, 244–263.

Tajima, K., Aminov, R. I., Nagamine, T., Matsui, H., Nakamura, M. et al. (2001) Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR. *Appl. Environ. Microbiol.*, **67**, 2766–2774.

Tiwari, P., Misra, B. N., and Sangwan, N. S. (2013) Beta-glucosidases from the fungus *Trichoderma*: an efficient cellulase machinery in biotechnological applications. *BioMed Res. Int.*, **2013**, e203735.

Wilson, D. B. (2008) Three microbial strategies for plant cell wall degradation. *Ann. N.Y. Acad. Sci.*, **1125**, 289–297.

Yang, B., Dai, Z., Ding, S. Y., and Wyman, C. E. (2014) Enzymatic hydrolysis of cellulosic biomass. *Biofuels*, **2**, 421–449.

You, M., Yue, Z., He, W., Yang, X., Yang, G. et al. (2013) A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.*, **45**, 220–225.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.*, **38**, e132.

Zhu, Z., Hang, S., Mao, S., and Zhu, W. (2014) Diversity of *Butyrivibrio* group bacteria in the rumen of goats and its response to the supplementation of garlic oil. *Asian-Australas. J. Anim. Sci.*, **27**, 179–186.