



A fast workflow to explore active enzymes from environmental samples through functional metagenomics

Arief Muammar^{1,2} · Endah Retnaningrum² · Budi Setiadi Daryono² · Irfan Dwidya Prijambada³ · Yuki Yashima^{1,4} · Clemens Peterbauer¹

Received: 24 December 2025 / Revised: 17 March 2026 / Accepted: 20 March 2026
© The Author(s) 2026

Abstract

Functional metagenomics has emerged as an effective tool for discovering novel enzymes directly from environmental samples, overcoming the limitations of traditional culture-based methods. In this study, we used a functional metagenomic approach on stool samples from *Axis kuhlii*, an endemic deer species from Indonesia, to identify active cellulases. We created an efficient workflow for expression of metagenomic sequences directly in *Komagatella phaffii* by combining metagenomic sequencing to investigate enzyme diversity, multiplex PCR to build a genes library, and rolling circle amplification (RCA) to streamline the cloning process, eliminating the need for intermediate *Escherichia coli* transformation and propagation steps. Furthermore, a semi-high-throughput screening method was used to evaluate multiple samples at once, allowing for the rapid identification of active enzymes. Using this approach, we discovered five endoglucanases and three β -glucosidases with confirmed enzyme activity. This study shows that functional metagenomics can bridge the gap between computational predictions and experimental validation, providing a reliable platform for enzyme discovery and characterization from complex environmental microbiomes.

Key points

- We established *K. phaffii* expression of metagenomic sequences via multiplex PCR and RCA.
- This approach links metagenomic and activity screening to enable enzyme discovery.
- Eight active cellulases were obtained from environmental samples through this approach.

Keywords Functional metagenomic · Multiplex PCR · Rolling circle amplification · Cellulase

Introduction

Metagenomics is a culture-independent approach that extracts and analyzes environmental DNA (eDNA) from microbial communities. This method is especially important because traditional microbiological techniques can only cultivate around 1% of prokaryotic microorganisms, leaving the vast majority of microbial diversity unexplored. By eliminating the need for cultivation, metagenomics enables researchers to access the genetic material of unculturable microorganisms, which account for roughly 99% of microbial diversity (Culligan et al. 2013). This metagenomic approach has transformed biotechnology by allowing the discovery of novel genes and enzymes with distinct properties.

Robinson et al. (2021) classified metagenomic enzyme discovery into three groups: De novo discovery: identifying entirely new enzyme types with no characterized members;

✉ Clemens Peterbauer
clemens.peterbauer@boku.ac.at

¹ Food Biotechnology Laboratory, Department of Biotechnology and Food Sciences, BOKU University, Muthgasse 11, 1190 Vienna, Austria

² Department of Tropical Biology, Faculty of Biology, Universitas Gadjah Mada, Jalan Teknik Selatan, Sinduadi, Mlati, Sleman 55281, Special Region of Yogyakarta, Indonesia

³ Faculty of Agriculture, Universitas Gadjah Mada, Jalan Flora, Bulaksumur, Sleman 55281, Special Region of Yogyakarta, Indonesia

⁴ Department of Biomaterial Sciences, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

Reference-based discovery: characterizing new reactions within known protein families; and Enzyme expansion: focusing on extending the properties of known enzyme classes. The term Enzyme expansion rather than enzyme discovery was used because this approach is mostly used to improve characterized enzymes with different substrate preferences. The de novo discovery represents the most challenging and novel discoveries, targeting entirely uncharacterized enzyme families. This approach is mostly used in cultivated microbes, to learn more about novel substrates, cofactors, and reactions. The reference-based discovery involves finding new functions within known families. This approach provides high-throughput annotation of eDNA datasets where some reference points are available.

Functional metagenomics, as a subset of metagenomic enzyme discovery, not only allows researchers to learn about the diversity and abundance of novel genes or enzymes, but also to obtain and produce functional novel proteins. Berini et al. (2017) reviewed 332 metagenome-sourced enzymes from 2014 to 2017, but only few have been optimized for industrial applications, leaving room for further development. The limited optimization of metagenome-sourced enzymes for industrial applications is due to technical, economic, and practical challenges, as well as a prioritization of discovery over application. This opens up significant opportunities for further development and industrial exploitation of these enzymes (Seo et al. 2014).

Lignocellulose has become a popular target for metagenomic enzyme discovery due to its abundance, complexity, and industrial potential. The metagenome sample we used was stool from an endemic and geographically isolated deer species, *Axis kuhlii*. It is native to the small island of Bawean in Indonesia and is classified as critically endangered. We compared the metagenomic sequences with known protein families (reference-based discovery) and used multiplex PCR and rolling circle amplification to prepare libraries specific for different cellulase activities. We then screened the libraries for endoglucanase and β -glucosidase activity, identifying several functional enzymes.

Material and methods

Cellulase diversity by metagenomic sequencing

Stool samples were collected from free-ranging *Axis kuhlii* from Bawean Island Wildlife Reserve (samples 1–3) as well as from captive animals in two different facilities within the wildlife reserve (samples 4–6 and 7–9). Metagenomic

DNA from the stool samples was extracted using the Quick-DNA™ HMW MagBead Kit from Zymo Research (Irvine, CA). A DNA library was prepared with the DNA library preparation kit and sequenced using a Next Generation Sequencing (NGS) approach on a Promethion sequencer (Oxford Nanopore Technologies, Oxford, UK). Basecalling was done using Guppy v2.2.3, and the raw reads were exported as fastq files and subsequently concatenated by sample. A quality control analysis of raw reads was performed using longQC.py v1.2.1 (Fukasawa et al. 2020), and the results were manually inspected. The raw reads were assembled using Canu v2.2 (parameters: -nanopore, genomeSize = 3.5 m) (Koren et al. 2017) and Flye v2.9.3 (parameters: --meta) (Kolmogorov et al. 2020). The assembled contigs were then error-polished using racon v1.5.0 with default parameters (Vaser et al. 2017).

Genes were predicted from the metagenome assemblies using Prodigal v2.6.3 (Hyatt et al. 2010). A clustering step was conducted using MMseqs2 v15.6f452 (Steinegger and Söding 2017). The gene catalog was annotated using a modification of the workflow available at <https://github.com/lucasz88/FunTraits> (Zoccarato et al. 2022). An HMM-based approach using KEGG Orthologs of endoglucanase (K01179) and β -glucosidase (K05349) was employed. The assembly and annotation of the gene catalog were carried out by the BOKU Core Facility Bioinformatics.

Amplification of cellulase genes by multiplex PCR

The most prominent cellulases from the eDNA (the same samples were used for NGS) were amplified using a multiplex PCR approach. We chose the annealing temperature (T_a) to be used around the central annealing temperature, and the selection process for the genes to be amplified was also based on the T_a being close, with a maximum difference of no more than 2 °C. Using this approach, we chose 16 sets of endoglucanase-specific primers and 10 sets of β -glucosidase-specific primers. Fifty microliters of reaction cocktails contained 0.2 mM dNTPs, 1× reaction buffer (New England Biolabs, Beverly, MA), 1U Q5 polymerase (New England Biolabs) and 150 nM of each primer. The amplicons were then inserted into the pPICZαA vector for *K. phaffii* X33 secretory expression (Invitrogen, Carlsbad, CA) with HiFi DNA assembly (New England Biolabs). NEBUILDER (v. 2.9.1; <https://nebbuilder.neb.com/#/>, accessed March 2025) was used to design primers. The list of primers used is given in Table S2. Coding sequences were cloned in a way that they contain a C-terminal His6 tag. The products obtained above were purified with the Monarch PCR & DNA Cleanup Kit (New England Biolabs). The GenBank accession numbers of the 16 endoglucanase and 10 β -glucosidase sequences are given in Table S3 in the Supplementary material.

Phi 29 rolling circle amplification

Obtained recombinant plasmids were transformed directly into *K. phaffii* without an amplification step in *E. coli* (Tachioka et al. 2016). To produce sufficient amounts of DNA for *K. phaffii* genomic integration, recombinant plasmid DNA needs to be amplified by Phi 29 polymerase (New England Biolabs) using RCA (Gibson et al. 2009; Hutchison et al. 2005) facilitated by random primer 6 (New England Biolabs). After 20 h of incubation at 30 °C, the mixture was heated to 65 °C for 10 min to inactivate the polymerase, and DNA amplification was confirmed by electrophoresis in a 0.8% agarose gel. The amplified DNA was then treated with *BlnI* (for endoglucanase) or *PmeI* (for β -glucosidase) (both from New England Biolabs) and a small aliquot was subjected to 0.8% agarose gel electrophoresis. The rest of the digested DNA was precipitated with ethanol for purification. At least 2 μ g linearized recombinant plasmid was transformed to *K. phaffii* using a standard electroporation protocol (Invitrogen) in a BioRad MicroPulser Electroporator (BioRad Laboratories, Hercules, CA). The transformed cells were plated on YPD agar with zeocin (100 μ g/ml) and incubated at 30 °C for three days.

96-well plate cultivation and screening

Single transformant colonies were transferred into deep-well microtiter plates containing 250 μ l of YPD media per well and incubated at 30 °C for 3 days with shaking at 350 rpm. Then, 250 μ l of YP medium was added. At 70, 82, and 106 h after inoculation, 50 μ l of methanol was added for induction of gene transcription via the AOX promoter. At 24 h after the last addition of inducer (for a total cultivation time of 130 h), the cells were harvested by centrifugation at 3500 rpm and 4 °C for 15 min, and the supernatants withdrawn for activity screening.

The supernatants were screened in 96-well plate format (Weis et al. 2004). Endoglucanase activity was determined using carboxymethyl cellulose as substrate and a Dinitrosalicylic Acid (DNS)-based assay (Miller 1959). Cellobiohydrolase activity was measured with 4-Methylumbelliferyl β -D-cellobioside (4-MUC), and 4-Methylumbelliferyl β -D-glucopyranoside (4-MUG) was used for β -glucosidase activity. These assays were adapted to microtiter plates: 80 μ l of the provided substrate solution were combined with 20 μ l sample (supernatant), mixed, and incubated at 50 °C for 15 min. For the DNS assay, 100 μ l DNS was added and the plate was heated in a boiling water bath for 15 min, then the absorbance at 540 nm was measured with a 2300 EnSpire Multilabel Plate Reader (PerkinElmer, Waltham, MA). For the 4-MUC and 4-MUG assays, absorbance was measured with excitation at 360 nm and an emission wavelength of 465 nm. Both assays were offset against a calibration curve generated with known standard concentrations (glucose standard concentrations for the DNS assay and 4-Methylumbelliferyl for the MUG and MUC-based assays).

Construct confirmation

Colonies presenting cellulase activity were confirmed by colony PCR. A single colony of each positive strain was used as DNA template for a 50 μ l PCR reaction, containing 0.2 mM dNTPs, 1 $\times$ reaction buffer, 1 U Q5 polymerase (all from New England Biolabs), 500 nM each of the pPICZ α A universal forward primer (5' TTAGCTTACTTTTCATAATTGC 3') and pPICZ α A universal reverse primer (5' AATGAAGCCTGC ATCTCTCAG 3'). The following program was used for PCR: 5 min at 98 °C, 35 cycles with 10 s at 98 °C, 30 s at 56 °C and 1.5 min at 72 °C, followed by a final extension step of 2 min at 72 °C. After gel electrophoresis, the identity of the resulting PCR products was verified by DNA Sanger sequencing.

SDS PAGE and Western Blot

Representatives of each colony with different active cellulases that were confirmed by sequencing were selected, the supernatants concentrated using Amicon Ultra 0.5 Centrifugal Filters (Sigma Aldrich, St. Louis, MO) and analyzed by SDS PAGE for endoglucanase and β -glucosidase using Mini-PROTEAN® TGX™ Precast Gels (BioRad Laboratories, Hercules, CA). Forty micrograms of protein was loaded per lane. For β -glucosidase, we also performed western blot using Penta-His Tag Monoclonal Antibody (Invitrogen, Carlsbad, CA) as primary antibody, and Polyclonal Rabbit Anti Mouse Immunoglobulin/HRP (Agilent Technologies, Santa Clara, CA) as secondary antibody. SDS PAGE units and blotting machine were from BioRad, visualization was done on a Gel Doc XR+ Gel Documentation System (BioRad Laboratories, Hercules, CA).

Sequence analysis

All cellulase genes identified through Sanger sequencing were analyzed for sequence similarity using BLAST search against the non-redundant (nr) GenBank, Uniprot/Swissprot, and Protein Data Base from NCBI. A phylogenetic tree was inferred using the Maximum Likelihood method implemented in the program MEGA12.1 (Kumar et al. 2024).

Results

Amplification of cellulase genes by multiplex PCR and RCA

We performed NGS on eDNA samples, then screened for and identified cellulase-encoding genes (Fig. 1). Metagenome data were obtained from nine stool samples of *Axis kuhlii* collected from the Nature Reserve and Wildlife Reserve on Bawean Island, Gresik, East Java, Indonesia;

31 endoglucanase and 154 β -glucosidase-encoding genes with different abundance distributions were identified in each sample (Fig. 2). We used the DNA from stool sample 5 for endoglucanase screening and stool sample 2 for β -glucosidase screening, because these samples had the highest cellulase distribution and abundance compared to the others.

For endoglucanases, from the 31 genes found in the sample, we selected genes with 5'- and 3'-termini allowing the design of oligonucleotide primers that had an annealing temperature in the range of 62–64 °C for multiplex PCR purposes, leaving 16 genes. For β -glucosidase, from the 154 genes found in the sample, we selected the ten most abundant genes and those representing different annotations in the sample, which had an annealing temperature for the multiplex primers of 64–66 °C (Table S1).

Multiplex PCR resulted in bands approximately 970–1000 bp in size with endoglucanase-specific primers, and approximately 2200–2300 bp in size with

β -glucosidase-specific primers. The amplicons were purified and ligated with linearized pPICZ α A with HIFI DNA Assembly to be used as plasmid template for rolling circle amplification. After cutting with *B*l*p*I (for endoglucanase) and *P*meI (for β -glucosidase), the DNA fragments from the RCA, which were initially larger than 10,000 bp (outside the separation range of the agarose gel), resulted in dominant DNA fragments of approximately 4400 bp for endoglucanase and approximately 5600 bp for β -glucosidase (Fig. S1). These sizes correspond to the pPICZ α A backbone (3400 bp) plus endoglucanase- or β -glucosidase-encoding genes.

Cultivation and screening for active enzymes in *K. phaffii* using 96-well plate format

After successful transformation into *K. phaffii*, more than 150 transformants for both endoglucanase- and β -glucosidase-encoding genes were cultivated in 96 plates for cellulase

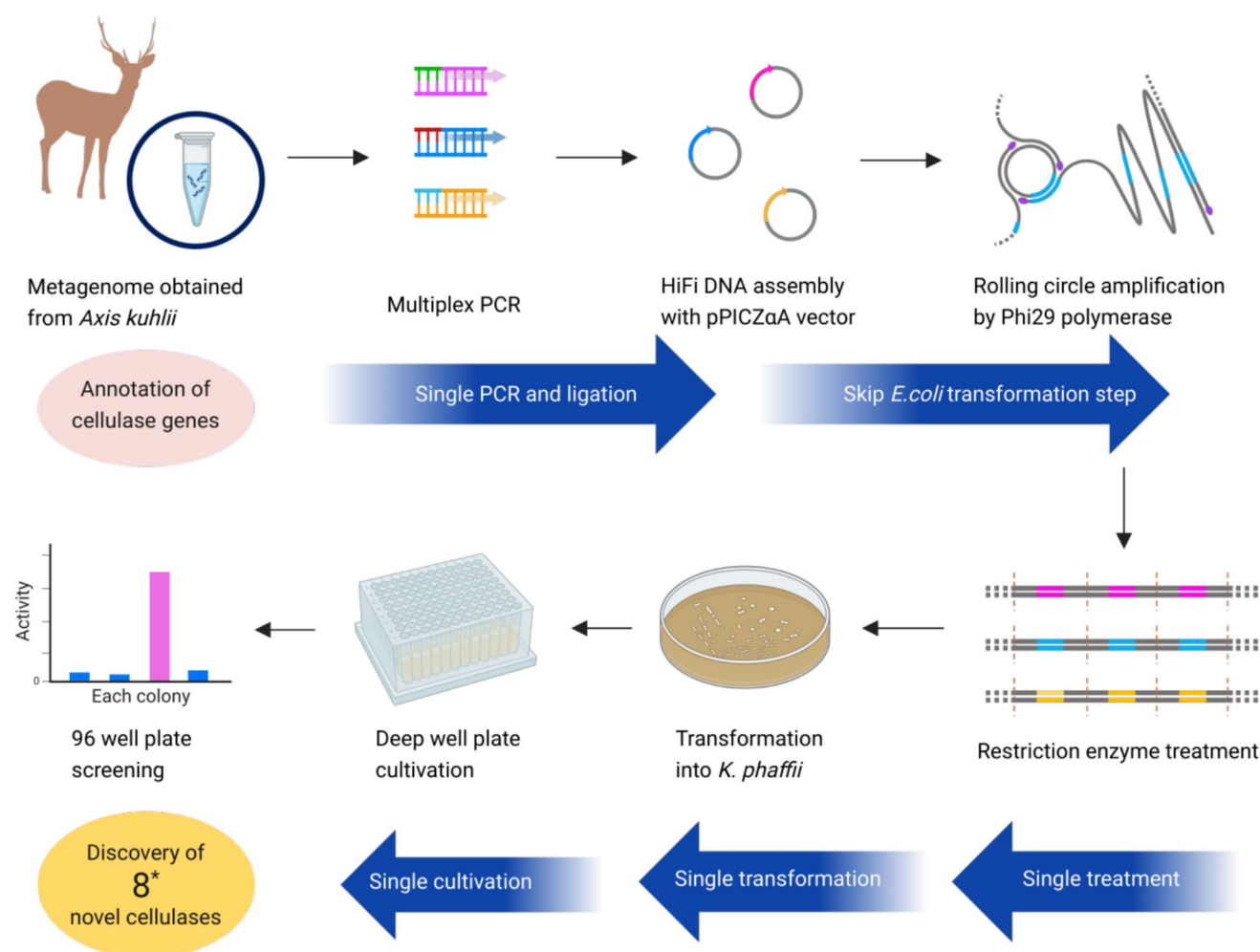
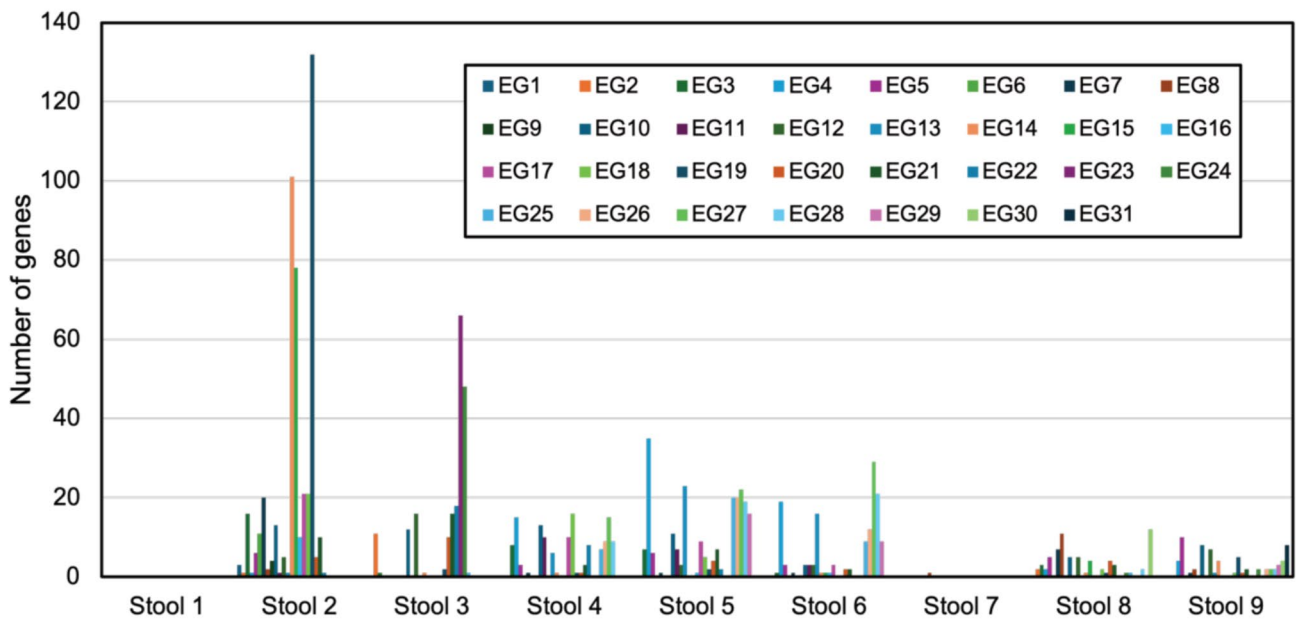


Fig. 1 Scheme of metagenomic strategies for the identification of novel cellulases from environmental samples. Figure created with Biorender.com and SILHOUETTE DESIGN. *We carried out this scheme for endoglucanases and β -glucosidases and found 8 novel cellulases in total

a



b

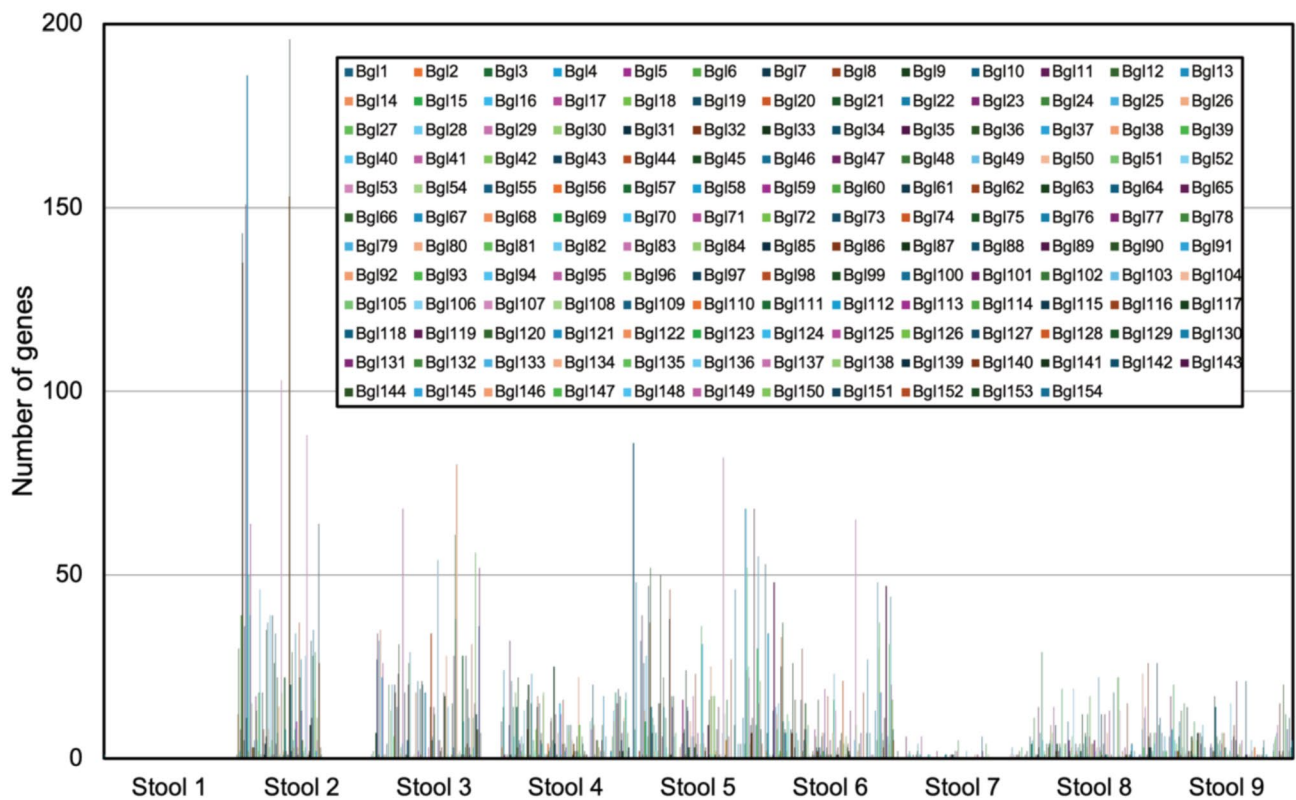


Fig. 2 Abundance of cellulases from 9 stool samples of *A. kuhlii*. **a** Endoglucanase and **b** β -glucosidase

activity screening. Colonies with highest activity above the threshold were selected for further testing and sequencing, namely 32 colonies containing an endoglucanase gene and 24 colonies containing a β -glucosidase gene (Fig. 3). The metagenome-derived inserts were identified by Sanger sequencing. 31 (out of 32) colonies harbored endoglucanase genes, and 11 (out of 24) harbored β -glucosidase genes.

Endoglucanase genes show a reasonably good ratio when comparing gene richness in the metagenomic samples to functional genes as transformants in *K. phaffii*. The number of DNA copies in the DNA template directly correlates with the number of transformants for each endoglucanase gene. There were 35 copies of the endoglucanase EG Q4 gene in the metagenome library, which produced 12 colonies of transformants with this gene. Similar to EG Q25, which had 20 gene copies in the DNA template, EG13 produced five colonies from 23 copies in the library. While the single transformant colony from seven gene copies of EG Q11 in the metagenomic template is below the otherwise observed ratio, eight transformant colonies for endoglucanase EG Q28 are somewhat numerous for the 19 templates. We also find that the number of copies in the metagenome template and the number of transformant colonies that contain the corresponding β -glucosidase genes are nearly equal. The ratio is significantly lower than that of the endoglucanase genes, though, as only eight colonies containing BGL1 were found out of the original 196 gene copies in the metagenome template. Only two colonies were obtained for BGL2, despite the abundance of 186 gene copies, whereas one transformant clone for BGL25 was produced from 30 gene copies, which is a roughly similar ratio (Table 1).

SDS PAGE of five endoglucanases and three β -glucosidases

The SDS PAGE results for endoglucanases show clear and dominant bands with a size ~ 34 kDa without prominent non-specific proteins, indicating optimal production of this protein in *K. phaffii*. Only in EG 11 the SDS PAGE showed a band indicating a smaller size (24 kDa instead of 34 kDa) in some, but not all cases. Expression of β -glucosidase genes was not as strong as that of endoglucanase genes, so the bands on SDS PAGE are not clearly visible. Therefore, Western blotting was performed to confirm the identity of these proteins (~ 83 kDa) (Fig. 4).

Rapid screening of colonies for other properties using a 96-well plate

We used a 96-deep well plate format to screen the obtained clones for activity with other substrates. Thirty-two colonies with high endoglucanase activity were screened for exoglucanase activity using 4MUC as a substrate (Fig. 5).

Thirteen clones were found to have exoglucanase activity, and all these contained endoglucanase genes 4, 13, 25, and 28. We also found one endoglucanase colony that could hydrolyse xylan (Fig. S2a), indicating that the endoglucanase it contained had promiscuous xylanase activity. Additionally, we could identify several colonies showing higher activity not only at standard pH and temperature conditions, but also at acidic and alkaline pH (pH 5.0 and 8.0, respectively) (Fig. S2b), as well as at elevated temperatures (60 °C) (Fig. S2c).

BLAST sequence comparisons

From the BLAST results with three different databases (non-redundant GenBank, Uniprot-Swissprot, and PDB), nr GenBank has the highest sequence query coverage and identity percentages because this database includes all sequences submitted to the database (such as NGS sequences, metagenome sequencing projects). In order to assess the novelty of the identified cellulase encoding genes, Uniprot-Swissprot and PDB are more appropriate (Table 2). The five endoglucanase genes that were successfully expressed showed similarity with sequences from the UniProtKB/SwissProt database below 50% (24% to 45%) and with sequences from the PDB database below 52% (between 41 and 51%). For β -glucosidase, the three enzymes we obtained showed sequence similarity with sequences from the UniProtKB/SwissProt database between 32 and 51%, and with the PDB database between 33 and 52%. From the eight sequences producing active enzymes and the ten best BLAST hits for each one (corrected for redundancy), a phylogenetic tree was inferred (Fig. 6).

Discussion

In this study, we successfully identified eight novel cellulases, five endoglucanases, and three β -glucosidases from the metagenome of the highly threatened Indonesian deer *Axis kuhlii*. By combining multiplex PCR of genes annotated as cellulases with RCA and functional screening in *K. phaffii*, functional enzymes were identified within 2 weeks after gene annotation. Conventional reference-based metagenomic screening approaches typically involve selecting a number of candidate genes and constructing individual expression strains one by one (Oliva et al. 2024; Paloyan et al. 2025). This workflow substantially reduced the required time and labor. The rapid identification of multiple functional cellulases highlights the efficiency of this strategy for enzyme discovery from eDNA.

The discovery of several cellulases using our method demonstrates that multiplex PCR using eDNA as a template

can work successfully and can be a cost-saving alternative compared to approaches using custom-synthesized genes. For successful multiplex PCR, primer-annealing temperature (T_a) interactions can impact multiplex performance (Lafrance et al. 2023). We designed primers with a maximum difference in T_a of no more than 2 °C and set the T_a for the annealing step around the central T_a for all primers. In addition to T_a , substantial differences in target sequence length may result in amplification bias, since longer templates require extended elongation times and are amplified less efficiently. A typical limitation of reference-based metagenomics is the difficulty to discover novel enzymes, which must rely on selecting target genes with relatively low similarity to known sequences. For example, an oxidative enzyme cleaving cellulose, representing a

previously unknown class compared with known enzymes, was found in the metagenome of long-term sugarcane bagasse-covered soil in Brazil by screening for candidate genes with a similarity of at least 10% identity with known carbohydrate-processing proteins (Santos et al. 2025). Although targeting genes with low sequence homology carries a higher risk of failure, multiplex PCR enables the simultaneous evaluation of both high- and low-risk candidates at once. Changing the portfolio of included target genes in multiplex PCR allows to balance between impact and likelihood of success. This strategy reduces experimental bias toward well-characterized enzymes and increases the likelihood of identifying novel functions.

Komagatella phaffii is a eukaryotic expression system that is capable of post-translational modifications and has

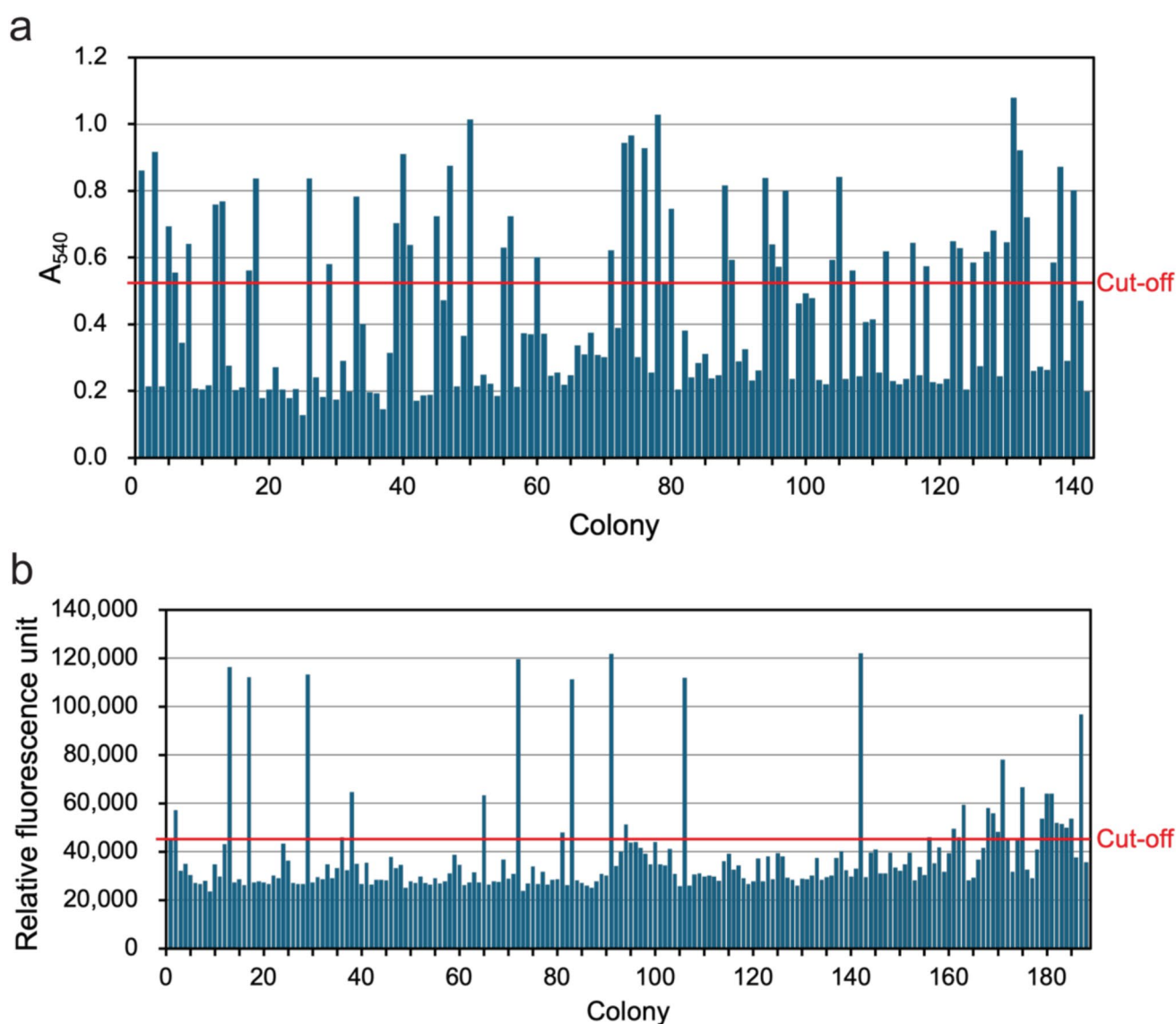


Fig. 3 Screening cellulases. **a** CMC assay for endoglucanase and **b** 4MUG assay for β -glucosidase. The last bar indicates the negative control

a powerful secretory system (Vijayakumar and Venkataraman 2023). However, transformation of *K. phaffii* generally requires several micrograms of plasmid DNA, which is substantially higher than that required for *E. coli*. This is generally done by transformation of *E. coli* after plasmid assembly, followed by cultivation and plasmid preparation. For the endoglucanase genes presented here, the conventional workflow requires 16 transformations into *E. coli* and *K. phaffii* individually. We applied rolling circle amplification (RCA) to generate sufficient amounts of plasmid DNA and transformed the reaction products digested by the restriction enzyme into *K. phaffii* at once. Successful RCA requires a completely ligated circular plasmid. Accordingly, ligation-based cloning methods, including HiFi DNA Assembly and Golden Gate Assembly, are compatible with RCA (Fryer et al. 2024), whereas seamless cloning systems that do not involve ligation, such as In-Fusion™ assembly (Zhu et al. 2007), are unsuitable. Overall, our approach can considerably shorten the process of expressing multiple genes derived from metagenomic samples in *K. phaffii*.

Functional screening was successful since the *K. phaffii* transformants contained various genes. The ratio of the transformants approximately reflected the numbers of the template gene copies in the metagenome sample in case of both endoglucanases and β -glucosidases. This suggests that multiplex PCR and RCA amplified different template genes almost equally for at least the positive target genes. Thus, the dominant gene in a template tends to be dominant in the population of transformants, although the ratio of all genes is effective for screening all target genes. Optimization of the primer concentration ratio (Lafrance et al. 2023) could be a solution to control the ratio of transformant genes. It should be noted that 180 transformants containing β -glucosidase genes proliferated, the majority of which might encode

β -glucosidases with no activity (7 out of 10 amplified genes). Only three genes (BGL 1, 2, and 25) were represented by eleven colonies containing active β -glucosidase. This also holds true for endoglucanase, where only five of sixteen endoglucanase genes produced clones with enzymatic activity. The reason for negative results can be a lack of amplification of the template genes, a failure of recombinant protein production or an inactive protein. The amplification process may favor genes that amplify easily, whereas genes with unfavourable CG-content or other problematic characteristics will produce fewer amplicons, which will be reflected in the quantity of intact expression plasmids. The generation of mini libraries through multiplex PCR amplification from eDNA as a template does not involve codon optimization (as custom gene synthesis would). This can result in a lower concentration of active enzyme in the supernatant, even if mRNA levels are comparable (Hanson and Collier 2018). Under unfavourable conditions, this may result in an assay readout that is near or below the detection limit, causing candidate genes to be missed. The complexity of using metagenomic DNA as a source of producing protein also can lead to difficulties in transcription and translation within hosts, often resulting in insufficient expression of foreign genes (Uchiyama and Miyazaki 2009). Even though we were able to obtain multiple functional enzymes using this method, this can be seen as a limitation.

The screening format in 96-well microtiter plates, which allows the processing of a greater number of constructs and screens for multiple properties (different substrates, optimal pH) in a standardized process, is another benefit of this strategy (Shang et al. 2018). For colonies transformed with endoglucanase genes, activity screening was performed not only using the DNS assay with CMC as a substrate, but also for cellobiohydrolase and xylanase activity, and optimal pH and thermostability screening could be performed as well. The majority of the active cellulase enzymes that we were able to obtain do not exhibit significant similarity to sequences found in the PDB and UniProtKB/SwissProt databases. BLAST results with UniProtKB/SwissProt and PDB show that our endoglucanases show a similarity with characterized enzymes between 24 and 51%. Similar functional metagenomic studies have obtained endoglucanase and xylanase genes from soil with 39% similarity (Chai et al. 2020), from Chinese yak rumen with 72% similarity (Chang et al. 2011), and from goat rumen with 56–59% similarity (Thapa et al. 2023). Our β -glucosidase samples showed similarities between 32 and 52% compared to 46–57% from kusaya gravy (Uchiyama et al. 2015), 87% from soil (Lu et al. 2013), and 52% from compost metagenome (Uchiyama et al. 2013). The BLAST results further indicate that most cellulases are bacterial enzymes. One β -glucosidase gene (BGL1) is more similar to eukaryotic sequences (β -glucosidase C from *Aspergillus*

Table 1 Diversity of cellulase genes in metagenome samples and transformants

	Metagenome sample		Transformants	
	Number of genes	Ratio (%)	Number of colonies	Ratio (%)
Endoglucanase				
EG4	35	34	12	39
EG11	7	7	1	3
EG13	23	22	5	16
EG25	20	19	5	16
EG28	19	18	8	26
Total	104	100	31	100
β -Glucosidase				
BGL 1	196	48	8	73
BGL 2	186	45	2	18
BGL 25	30	7	1	9
Total	412	100	11	100

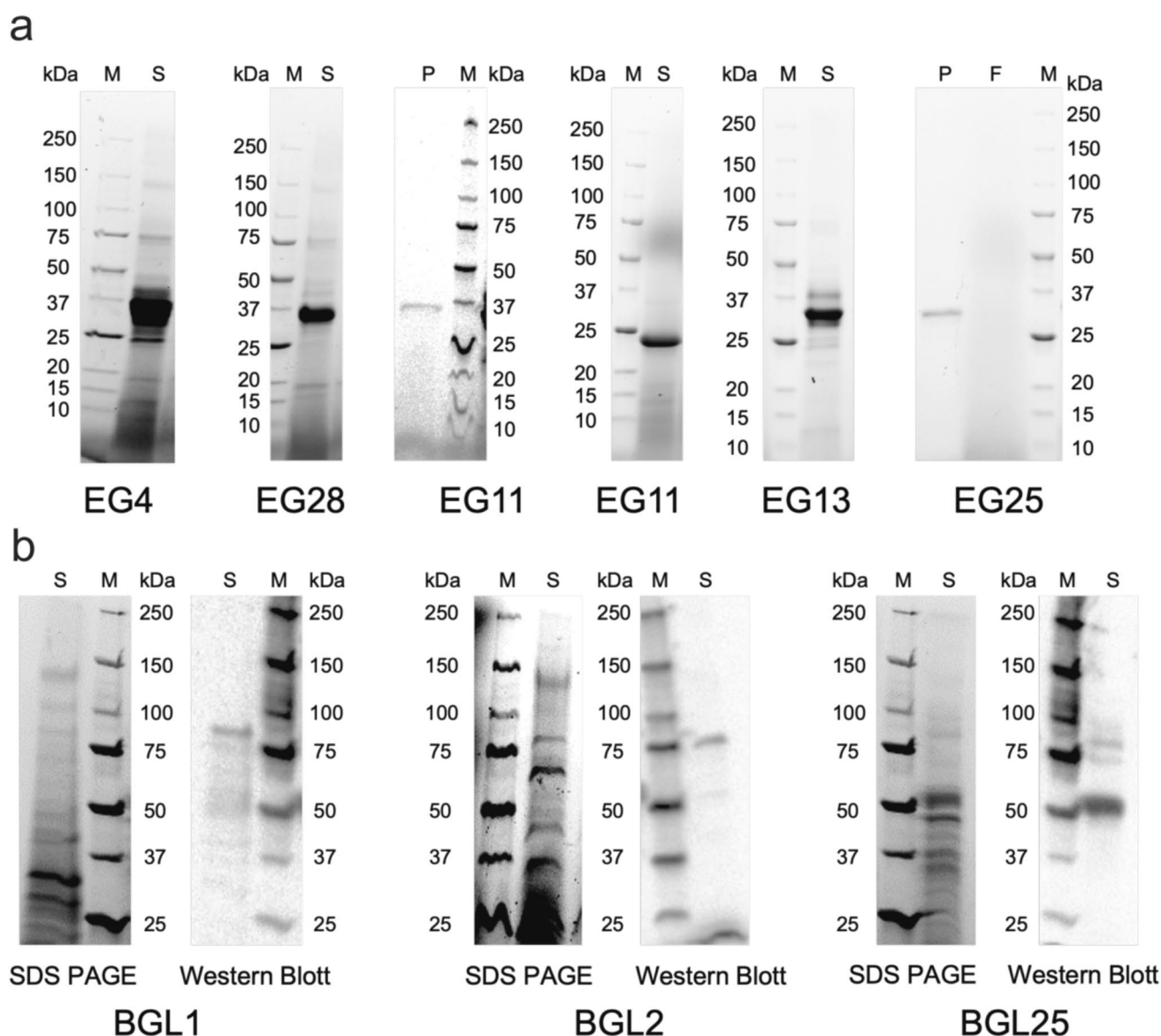


Fig. 4 **a** SDS PAGE for endoglucanase and **b** SDS PAGE and Western blot for β -glucosidase. M, marker; S, supernatant; P, pure; F, flow-through

flavus). *K. phaffii* is generally suited for eukaryotic genes, especially genes encoding secretory proteins that require post-translational modifications, but it can also express many prokaryotic genes (Lv and Cai 2025). These and our results demonstrate that the use of *K. phaffii* can lead to successful expression of metagenomic genes and yield functional enzymes even when the metagenome samples are assumed to be dominated by prokaryotic DNA.

Komagatella phaffii is a well-established host platform for the secretory production of recombinant proteins (Weis et al. 2004). If a gene found in this way satisfies specific requirements, positive clones could be transferred to process optimization and upscaling quickly. With a few minor

modifications, like the recircularization procedure prior to transformation, the workflow can be transferred to other host systems, such as *Bacillus subtilis*-based platforms. In order to maintain the expression construct as a circular, autonomously replicating plasmid rather than integrated into the genome as is required for *K. phaffii*, a recirculation with ligase is necessary because the DNA fragment obtained prior to transformation is a linear fragment. The same sets of substrate analogues for fluorescent detection can be used for secretory enzymes; for non-secretory enzymes an adaptation to *E. coli* is feasible as well, if an appropriate screening assay is available.

Fig. 5 Screening exoglucanase (cellobiohydrolase) activity using 4MUC in endoglucanase transformants. The last bar indicates the negative control

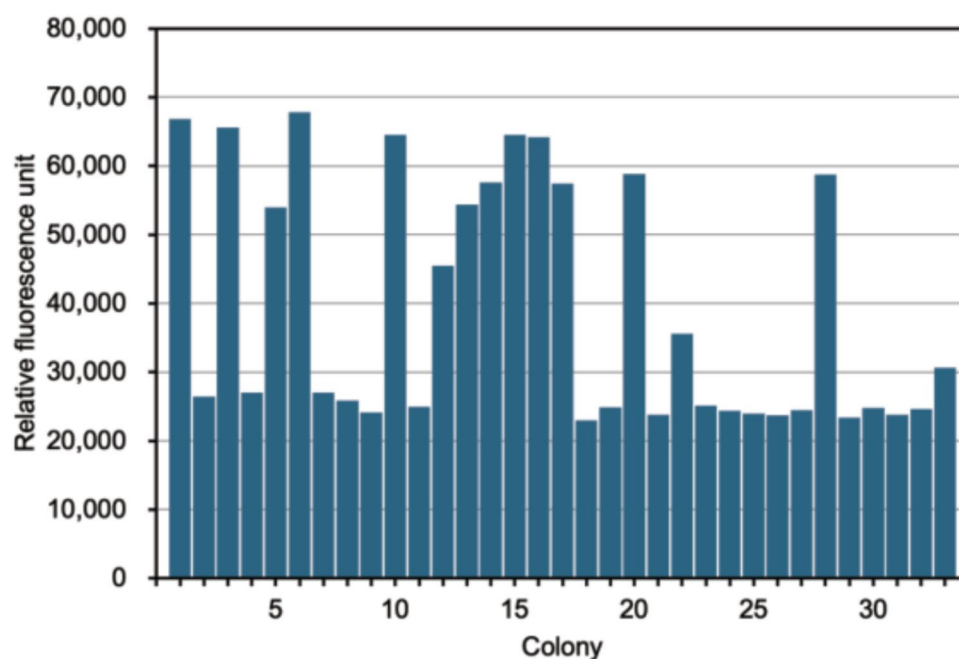


Table 2 Blast results with nr GenBank, Uniprot-Swissprot, and PDB databases

No.	Query	Database	Description	Organism	Cov (%)	ID (%)	Accession
Endoglucanase							
1	EG Q4	GenBank	CAZy GH 5	<i>Bacteroidales bacterium</i>	99	88	MBR2064430.1
		UniProtKB-SwossProt	Endoglucanase Z, Cellulase Z	<i>Dickeya dadantii</i>	87	45	P07103.2
		PDB	Chain A, Endoglucanase	<i>Cytophaga hutchinsonii</i>	100	42	5IHS_A
2	EG Q13	GenBank	CAZy GH 5	<i>Bacteroidales bacterium</i>	96	98	MBE6235115.1
		UniProtKB-SwossProt	Endoglucanase Z, Cellulase Z	<i>Dickeya dadantii</i>	92	43	P07103.2
		PDB	Endoglucanase	<i>Cellvibrio japonicus</i>	92	45	8BQA_AAA
3	EG Q25	GenBank	CAZy GH 5	<i>Bacteroidales bacterium</i>	99	100	MBE6235412.1
		UniProtKB-SwossProt	Endoglucanase Z, Cellulase Z	<i>Butyrivibrio fibrisolvens</i>	72	24	P22541.1
		PDB	Cellulase, putative, cel5C	<i>Cellvibrio japonicus</i>	91	51	8BN7_AAA
4	EG Q28	GenBank	Cellulase family GH	<i>Bacteroidales bacterium</i>	99	88	MBO5016517.1
		UniProtKB-SwossProt	Endoglucanase A, Cellulase A	<i>Butyrivibrio fibrisolvens</i>	62	26	P22541.1
		PDB	Cellulase, putative, cel5C	<i>Cellvibrio japonicus</i>	91	49	8BN7_AAA
5	EG Q11	GenBank	CAZy GH 5	<i>Bacteroidales bacterium</i>	99	84	MEE0236234.1
		UniProtKB-SwossProt	Endoglucanase Z, Cellulase Z	<i>Dickeya dadantii</i>	92	42	P07103.2
		PDB	Endoglucanase, CAZy GH 5	<i>Cytophaga hutchinsonii</i>	100	41	5IHS_A
β-Glucosidase (BGL)							
1	BGL 1	GenBank	CAZy GH 3	<i>Bacteroidales bacterium</i>	100	99	MBQ3251508.1
		UniProtKB-SwossProt	Probble β-glucosidase C	<i>Aspergillus flavus</i>	90	32	B8NGU6.1
		PDB	β-Glucosidse	<i>Clavibacter michiganensis</i>	92	33	9UL7_A
2	BGL 2	GenBank	CAZy GH 3	<i>Bacteroidales bacterium</i>	100	99	MBQ8810697.1
		UniProtKB-SwossProt	Thermostble β-glucosidase B	<i>Acetivibrio thermocellus</i>	92	37	P14002.2
		PDB	β-Glucosidase	<i>Segatella copri</i>	93	52	8IDZ_A
3	BGL 25	GenBank	CAZy GH 3	<i>Bacteroidales bacterium</i>	100	100	MBO5108178.1
		UniProtKB-SwossProt	β-Glucosidse CAZy GH 3	<i>Bacteroides ovatus</i> ATCC 8483	100	51	A7LXS8.1
		PDB	Thermostble β-glucosidase B	<i>Acetivibrio thermocellus</i> AD2	95	41	7MS2_A

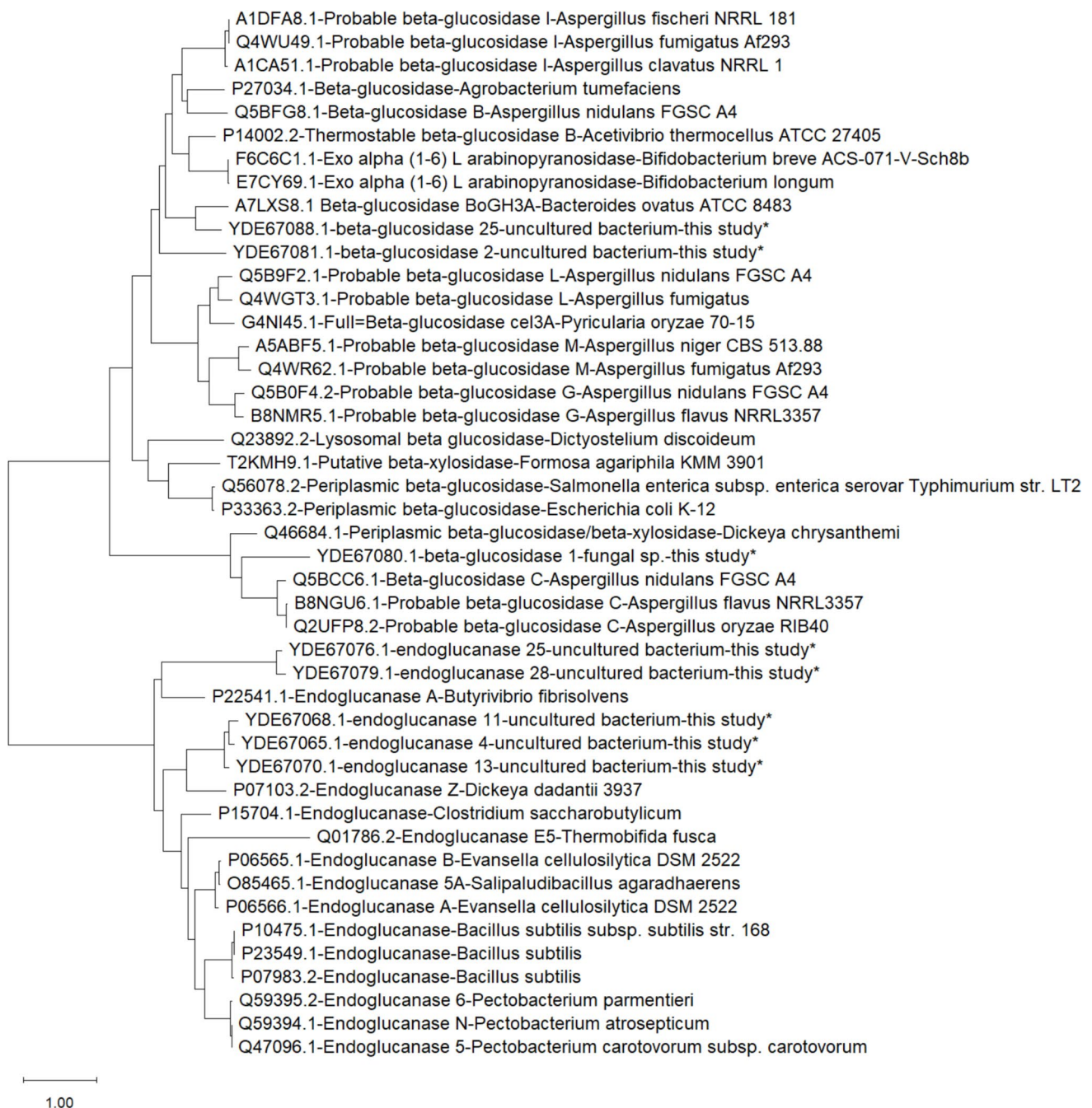


Fig. 6 Phylogenetic tree inferred from the eight sequences resulting in active enzymes and the ten best BLAST hits for each one. The sequences described in this study are marked with an asterisk

Conclusion

With the approach presented in this study, which combines metagenomics sequencing, multiplex PCR, rolling circle amplification, and the use of semi-high-throughput assays with deep well plates for rapid screening, we were able to identify novel variants of two different enzyme families

encoded in environmental samples more quickly and cost-efficiently. With samples sourced from the environment, this approach covers all genes represented in microorganisms, including those that cannot be cultivated. This approach has proven effective for two different enzyme groups, namely endoglucanase and β -glucosidase, and can be applied for other enzyme groups as well.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00253-026-13805-1>.

Acknowledgements The authors thank the Center for Higher Education Funding and the Indonesia Endowment Fund for Education (Scholarship approval letter number: 1688/J5.2.3./BPI.06/10/2021) for AM. YY was a beneficiary of the Overseas Challenge Program for Young Researchers from the Japan Society for the Promotion of Science (JSPS).

Author contributions Conceptualization: AM, CP, ER, BSD and IDP. AM, YY and CP designed research. AM conducted experiments. AM, YY and CP analyzed data. The first draft of the manuscript was written by AM, YY and CP, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by University of Natural Resources and Life Sciences Vienna (BOKU).

Data availability The data are given in the text and the supplemented files.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Berini F, Casciello C, Marcone GL, Marinelli F (2017) Metagenomics: novel enzymes from non-culturable microbes. *FEMS Microbiol Lett* 364:fnx211. <https://doi.org/10.1093/femsle/fnx211>
- Chai S, Zhang X, Jia Z, Xu X, Zhang Y, Wang S, Feng Z (2020) Identification and characterization of a novel bifunctional cellulase/hemicellulase from a soil metagenomic library. *Appl Microbiol Biotechnol* 104:7563–7572. <https://doi.org/10.1007/s00253-020-10766-x>
- Chang L, Ding M, Bao L, Chen Y, Zhou J, Lu H (2011) Characterization of a bifunctional xylanase/endoglucanase from yak rumen microorganisms. *Appl Microbiol Biotechnol* 90:1933–1942. <https://doi.org/10.1007/s00253-011-3182-x>
- Culligan EP, Sleator RD, Marchesi JR, Hill C (2013) Metagenomics and novel gene discovery: promise and potential for novel therapeutics. *Virulence* 5(3):399–412. <https://doi.org/10.4161/viru.27208>
- Fryer T, Wolff DS, Overath SE, Laustsen AH, Jenkins TP, Andersen C (2025) Post-assembly plasmid amplification for increased transformation yields in *E. coli* and *S. cerevisiae*. *Chem Bio Eng* 2:87–96. <https://doi.org/10.1021/cbe.4c00115>
- Fukasawa Y, Ermini L, Wang H, Carty K, Cheung M-S (2020) LongQC: a quality control tool for third generation sequencing long read data. *G3 Genes|genomes|genetics* 10:1193–1196. <https://doi.org/10.1534/g3.119.400864>
- Gibson D, Young L, Chuang R-Y, Venter JC, Hutchison CA III, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6:343–345. <https://doi.org/10.1038/nmeth.1318>
- Hanson G, Collier J (2018) Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* 19:20–30. <https://doi.org/10.1038/nrm.2017.91>
- Hutchison CA, Smith HO, Pfannkuch C, Venter JC (2005) Cell-free cloning using ϕ 29 DNA polymerase. *Proc Natl Acad Sci U S A* 102(48):17332–17336. <https://doi.org/10.1073/pnas.0508809102>
- Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. <https://doi.org/10.1186/1471-2105-11-119>
- Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA (2020) MetaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 17:1103–1110. <https://doi.org/10.1038/s41592-020-00971-x>
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 27:722–736. <https://doi.org/10.1101/gr.215087.116>
- Kumar S, Stecher G, Suleski M, Sanderford M, Sharma S, Tamura K (2024) MEGA 12: molecular evolutionary genetics analysis version 12 for adaptive and green computing. *Mol Biol Evol* 41:1–9. <https://doi.org/10.1093/molbev/msae263>
- Lafrance R, Valdez-Torres JB, Villicaña C, García-Estrada RS, Esparza-Araiza MJ, León-Félix J (2023) Response surface methodology for optimization of multiplex-PCR protocols for detection of TYLCV, TSWV and Fol molecular markers: analytical performance evaluation. *Genes* 14:337. <https://doi.org/10.3390/genes14020337>
- Lu J, Du L, Wei Y et al (2013) Expression and characterization of a novel highly glucose-tolerant β -glucosidase from a soil metagenome. *Acta Biochim Biophys Sin* 45:664–673. <https://doi.org/10.1093/abbs/gmt061>
- Lv W, Cai M (2025) Advancing recombinant protein expression in *Komagataella phaffii*: opportunities and challenges. *FEMS Yeast Res* 25:foaf010. <https://doi.org/10.1093/femsyr/foaf010>
- Miller GL (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal Chem* 31:426–428. <https://doi.org/10.1021/ac60147a030>
- Oliva B, Zervas A, Stougaard P, Westh P, Thøgersen MS (2024) Metagenomic exploration of cold-active enzymes for detergent applications: characterization of a novel, cold-active and alkali-stable GH8 endoglucanase from ikaite columns in SW Greenland. *Microb Biotechnol* 17:e14466. <https://doi.org/10.1111/1751-7915.14466>
- Paloyan A, Soghomonyan T, Karapetyan M, Grigoryan H, Krüger A, Cuskin F, Marles-Wright J, Burkhardt C, Antranikian G (2025) A novel acidic laminarinase derived from Jermuk hot spring metagenome. *Appl Microbiol Biotechnol* 109:172. <https://doi.org/10.1007/s00253-025-13557-4>
- Robinson SL, Piel J, Sunagawa S (2021) A roadmap for metagenomic enzyme discovery. *Nat Prod Rep* 38(11):1994–2023. <https://doi.org/10.1039/d1np00006c>
- Santos CA, Morais MAB, Mandelli F, Lima EA, Miyamoto RY, Higasi PMR, Araujo EA, Paixão DAA, Junior JM, Motta ML, Streit RSA, Morão LG, Silva CBC, Wolf LD, Terrasan CRF, Bulka NR, Diogo JA, Fuzita FJ, Colombari FM, Santos CR, Rodrigues PT, Silva DB, Grisel S, Bernardes JS, Terrapon N, Lombard V, Filho AJC, Henrissat B, Bissaro B, Berrin JG, Persinoti GF, Murakami MT (2025) A metagenomic ‘dark matter’ enzyme catalyses oxidative cellulose conversion. *Nature* 639:1076–1083. <https://doi.org/10.1038/s41586-024-08553-z>

- Seo S, Lee YS, Yoon SH (2014) Characterization of a novel cold-active esterase isolated from swamp sediment metagenome. *World J Microbiol Biotechnol* 30:879–886. <https://doi.org/10.1007/s11274-013-1496-9>
- Shang M, Chan VJ, Wong DWS, Liao H (2018) A novel method for rapid and sensitive metagenomic activity screening. *MethodsX* 5:669–675. <https://doi.org/10.1016/j.mex.2018.06.011>
- Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026–1028. <https://doi.org/10.1038/nbt.3988>
- Tachioka M, Sugimoto N, Nakamura A, Sunagawa N, Ishida T, Uchiyama T, Igarashi K, Samejima M (2016) Development of simple random mutagenesis protocol for the protein expression system in *Pichia pastoris*. *Biotechnol Biofuels* 9(1):1–10. <https://doi.org/10.1186/s13068-016-0613-z>
- Thapa S, Zhou S, O'Hair J, Nasr KA, Ropelewski A, Li H (2023) Exploring the microbial diversity and characterization of cellulase and hemicellulase genes in goat rumen: a metagenomic approach. *BMC Biotechnol* 23:51. <https://doi.org/10.1186/s12896-023-00821-6>
- Uchiyama T, Miyazaki K (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr Opin Biotechnol* 20(6):616–622. <https://doi.org/10.1016/j.copbio.2009.09.010>
- Uchiyama T, Miyazaki K, Yaoi K (2013) Characterization of a novel β -glucosidase from a compost microbial metagenome with strong transglycosylation activity. *J Biol Chem* 288(25):18325–18334. <https://doi.org/10.1074/jbc.M113.471342>
- Uchiyama T, Yaoi K, Miyazaki K (2015) Glucose-tolerant β -glucosidase retrieved from a Kusaya gravity metagenome. *Front Microbiol* 6:548. <https://doi.org/10.3389/fmicb.2015.00548>
- Vaser R, Sović I, Nagarajan N, Šikić M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27:737–746. <https://doi.org/10.1101/gr.214270.116>
- Vijayakumar V, Venkataraman K (2023) A systematic review of the potential of *Pichia pastoris* (*Komagataella phaffii*) as an alternative host for biologics production. *Mol Biotechnol* 66:1621–1639. <https://doi.org/10.1007/s12033-023-00803-1>
- Weis R, Luiten R, Skranc W, Schwab H, Wubbolts M, Glieder A (2004) Reliable high-throughput screening with *Pichia pastoris* by limiting yeast cell death phenomena. *FEMS Yeast Res* 5(2):179–189. <https://doi.org/10.1016/j.femsyr.2004.06.016>
- Zhu B, Cai G, Hall EO, Freeman GJ (2007) In-FusionTM assembly: seamless engineering of multidomain fusion proteins, modular vectors, and mutations. *Biotechniques* 43:354–359. <https://doi.org/10.2144/000112536>
- Zoccarato L, Sher D, Miki T, Segrè D, Grossart H-P (2022) A comparative whole-genome approach identifies bacterial traits for marine microbial interactions. *Commun Biol* 5:276. <https://doi.org/10.1038/s42003-022-03184-4>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.