VSOB-4 Bioinformatics course:
Metagenome analysis in one health practice – from 16S to shotgun

Metagenomic approaches in studying biodiversity of bacterial communities and mining lignocellulose-degrading genes from metagenomic DNA data

Prof.Dr. Truong Nam Hai
Institute of Biology (IB)
Vietnam Academy of Science and Technology (VAST)

Quy Nhon, September 3, 2025

CONTENTS

1. Introduction

Why *meta*GENOMICS?

- **Powerful technology**
  - Analyzes microbial communities via metagenomes, not individual genomes.
- **System-level approach**
  - Reveals biological mechanisms in microbial ecosystems.
  - Supports solutions to complex biological problems.
- **Culture-independent**
  - No need to grow each species separately.
  - Works directly from soil, water, gut, etc.
- **Critical advantage**
  - ~99% of microbes cannot be cultured by conventional methods.

What can metagenomics do?

- **Analyze Microbial–Environment Interactions**
  Understand how microbial communities respond to environmental factors such as pH, temperature, salinity, and nutrient availability.
- **Study Microbial–Host Relationships**
  Explore how microbiomes influence host health, disease, immunity, and metabolic functions.
- **Discover High-Value Biomolecules**
  Identify novel enzymes, secondary metabolites, and other functional biomolecules with potential applications in **biotechnology**, **agriculture**, and **medicine**.
- **Support Environmental Monitoring and Bioremediation**
  Track ecosystem changes and identify microbes capable of degrading pollutants or improving soil and water quality.
- **Enable Functional and Ecological Insights**
  Reveal the structure, dynamics, and functions of microbial communities in complex ecosystems - without the need for culturing.

How are target genes identified in metagenomics?

❖Functional Screening:

- **Based on protein and enzyme functions**
  Identify genes encoding enzymes or functional proteins of interest.
- **Approach:**
  - Construct metagenomic libraries (e.g., using plasmids or fosmids) by cloning environmental DNA into expression hosts (such as *E. coli*).
  - Screen the library for desired activities (e.g., cellulase, lipase, protease) using functional assays.
  - Positive clones are then sequenced to identify novel genes and enzymes.

❖Sequence-Based Screening (using NGS data):

- **16S rRNA sequencing**
  Taxonomic identification and profiling of microbial communities.
- **Whole-metagenome shotgun sequencing**
  Comprehensive analysis of taxonomic composition and functional potential proteins or enzymes.
- **Bioinformatics tools**
  Analyze and filter large-scale sequence data to identify organisms or genes of interest.

Significance of lignocellulose

~40 million tons of rice straws/year in Vietnam

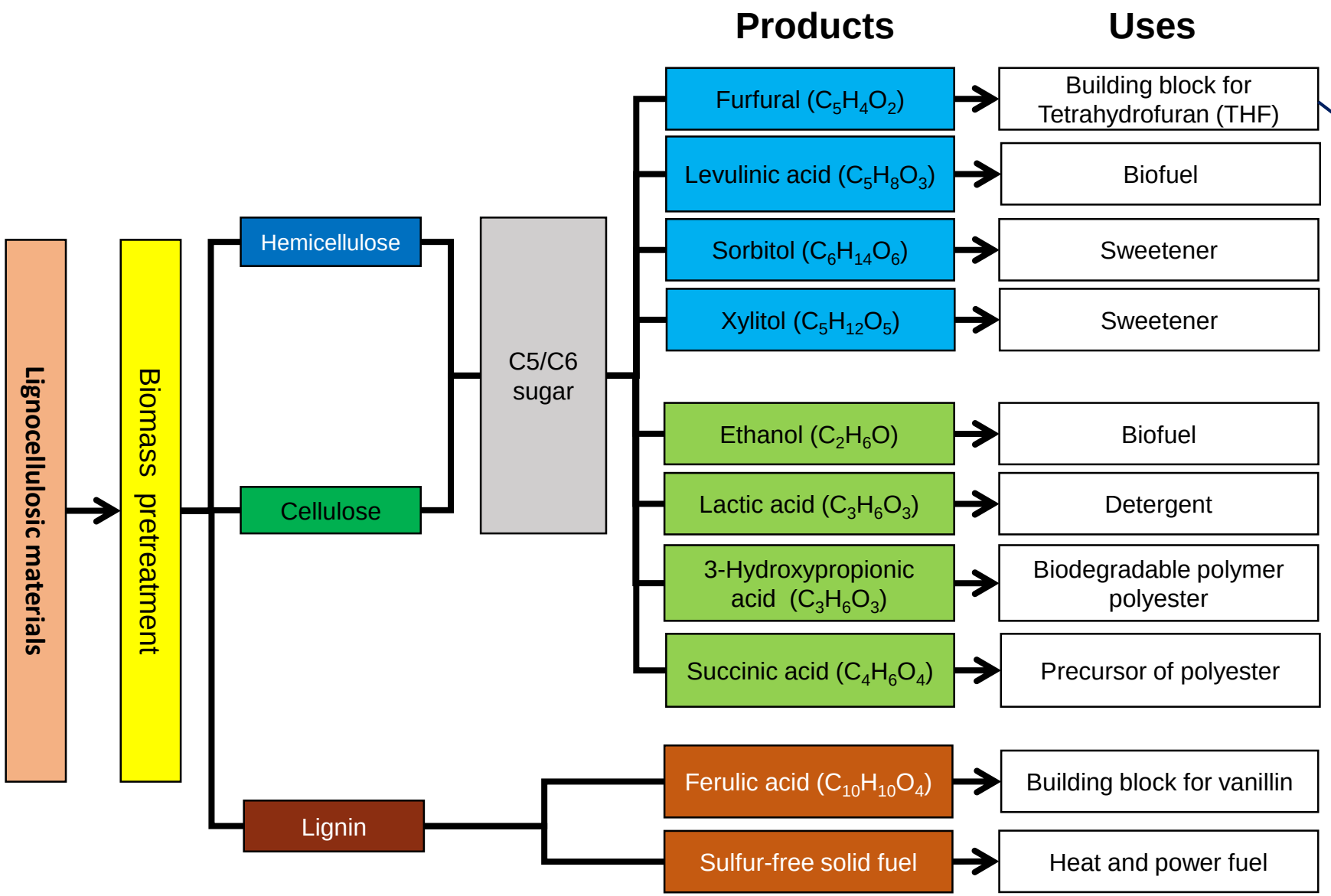
Animal feed (supplement for ruminants)

Mushroom cultivation (e.g., oyster mushrooms)

Burning (major cause of air pollution)

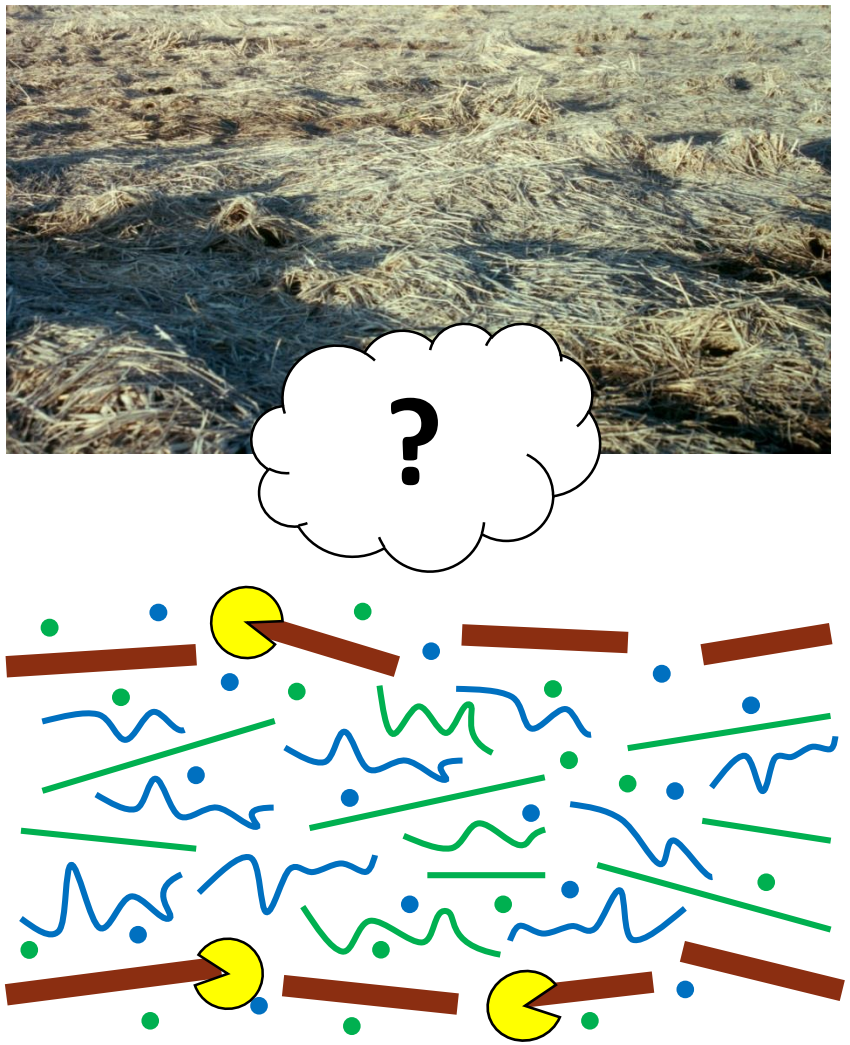
Effective utilization can reduce waste and environmental impact.

Recycle of Lignocellulose components



- Solvents in laboratories and industry;
- Applications in polymer production;
- Use in the chemical industry;
- Solvents in paints, printing inks, and adhesives;
- In the pharmaceutical industry.

Natural decomposition

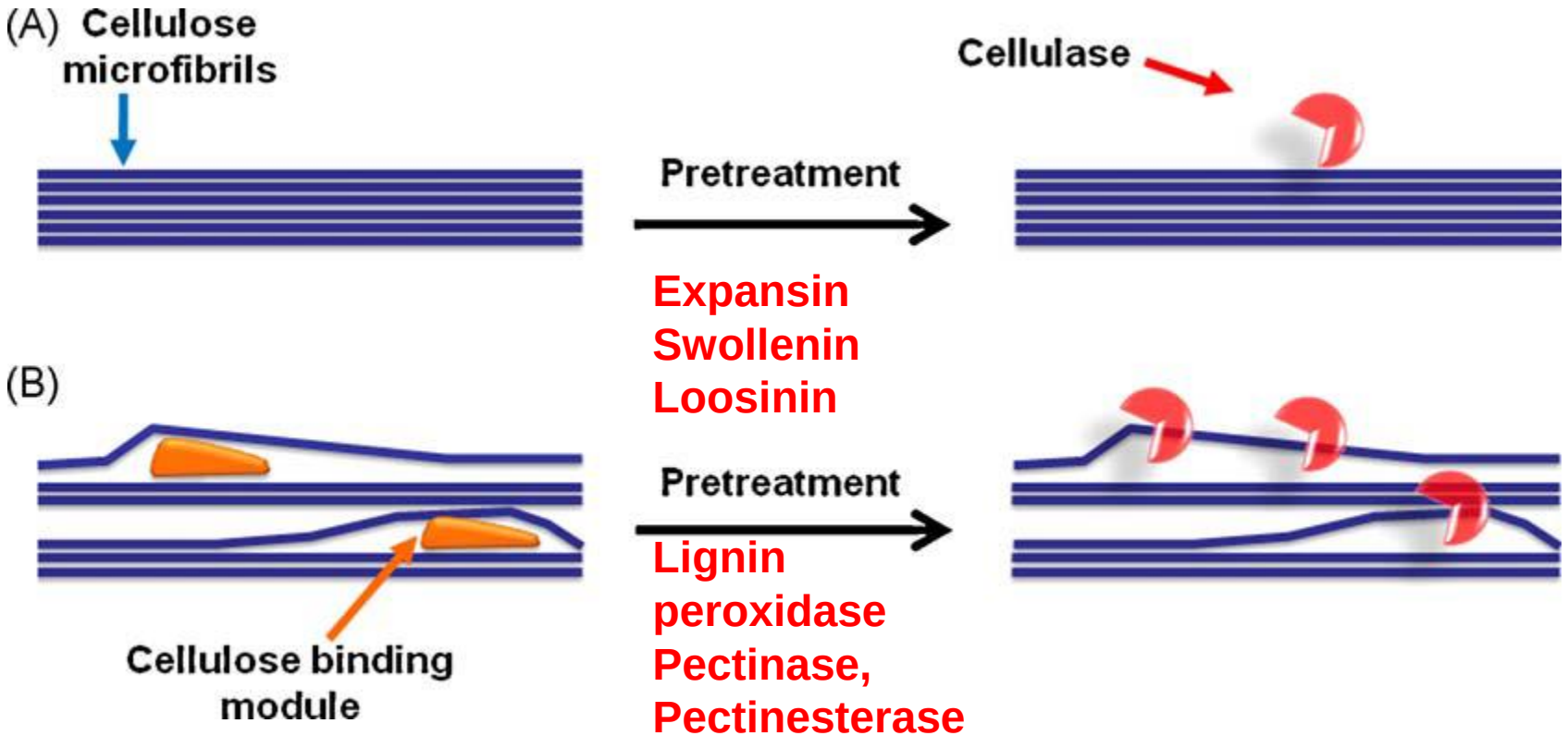


Lignin

Cellulose

Hemicellulose

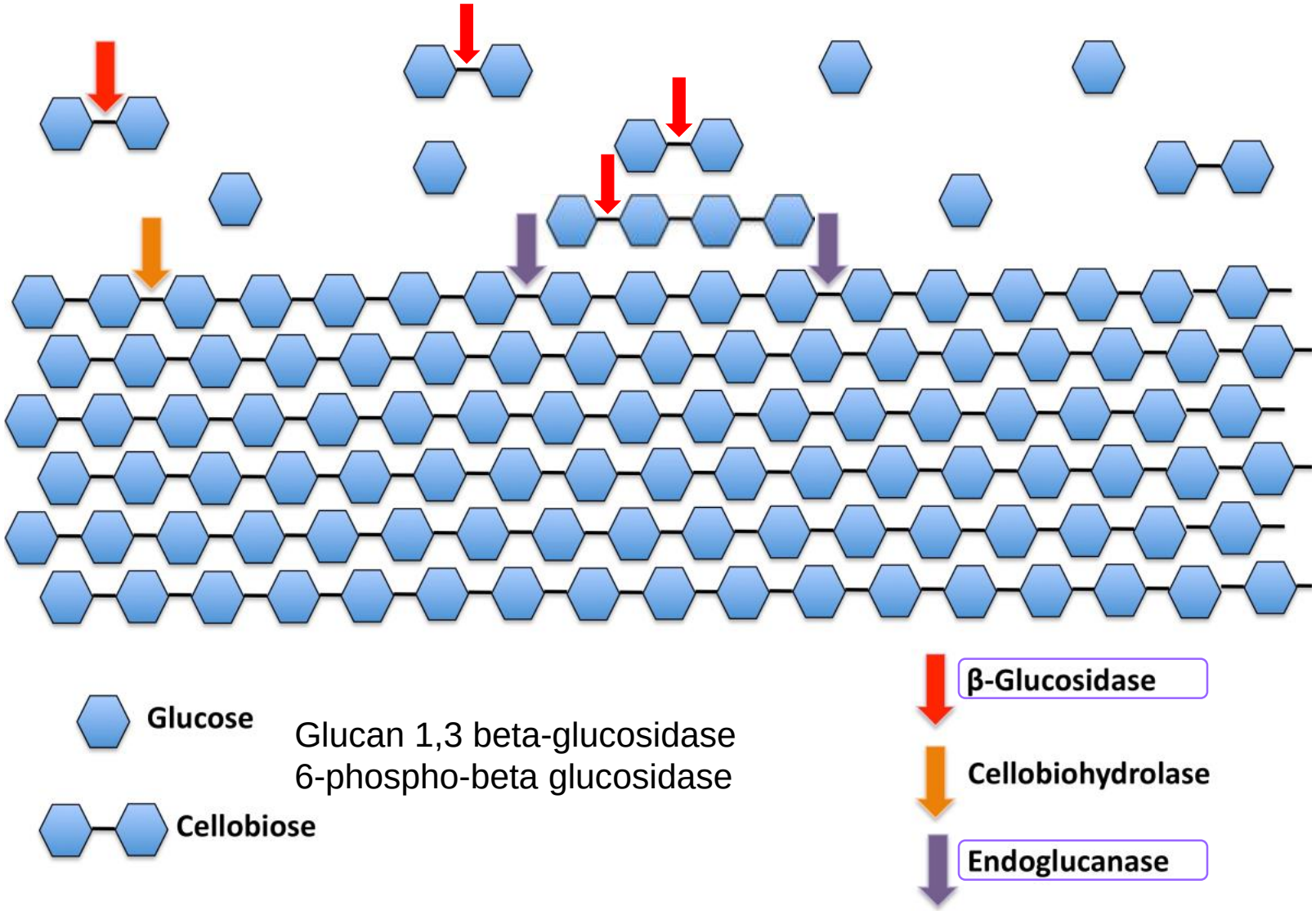
Enzymes and proteins involving in natural pretreatment



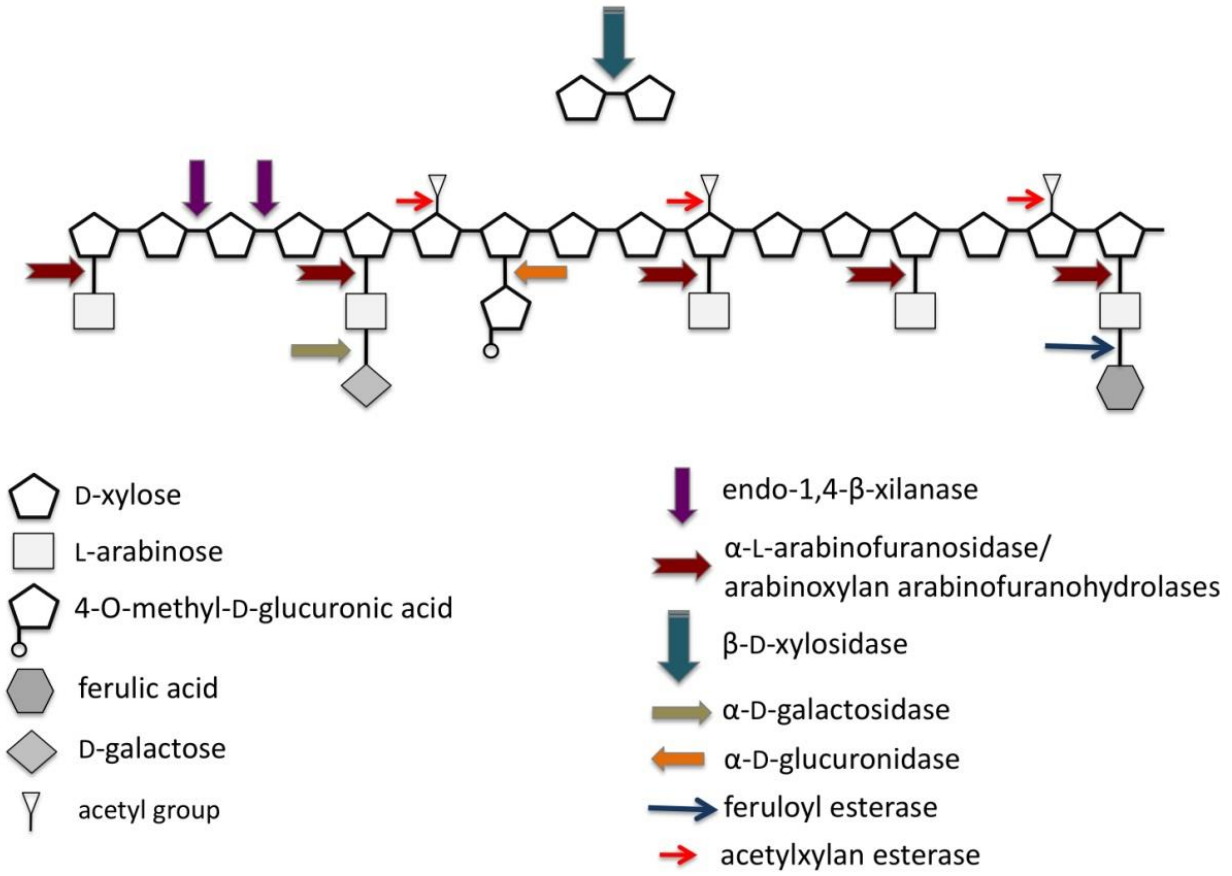
Expansin
Swollenin
Loosinin

Lignin peroxidase
Pectinase, Pectinesterase
Peroxidase
LPMOs
Lacase
…

Cellulose degrading enzymes



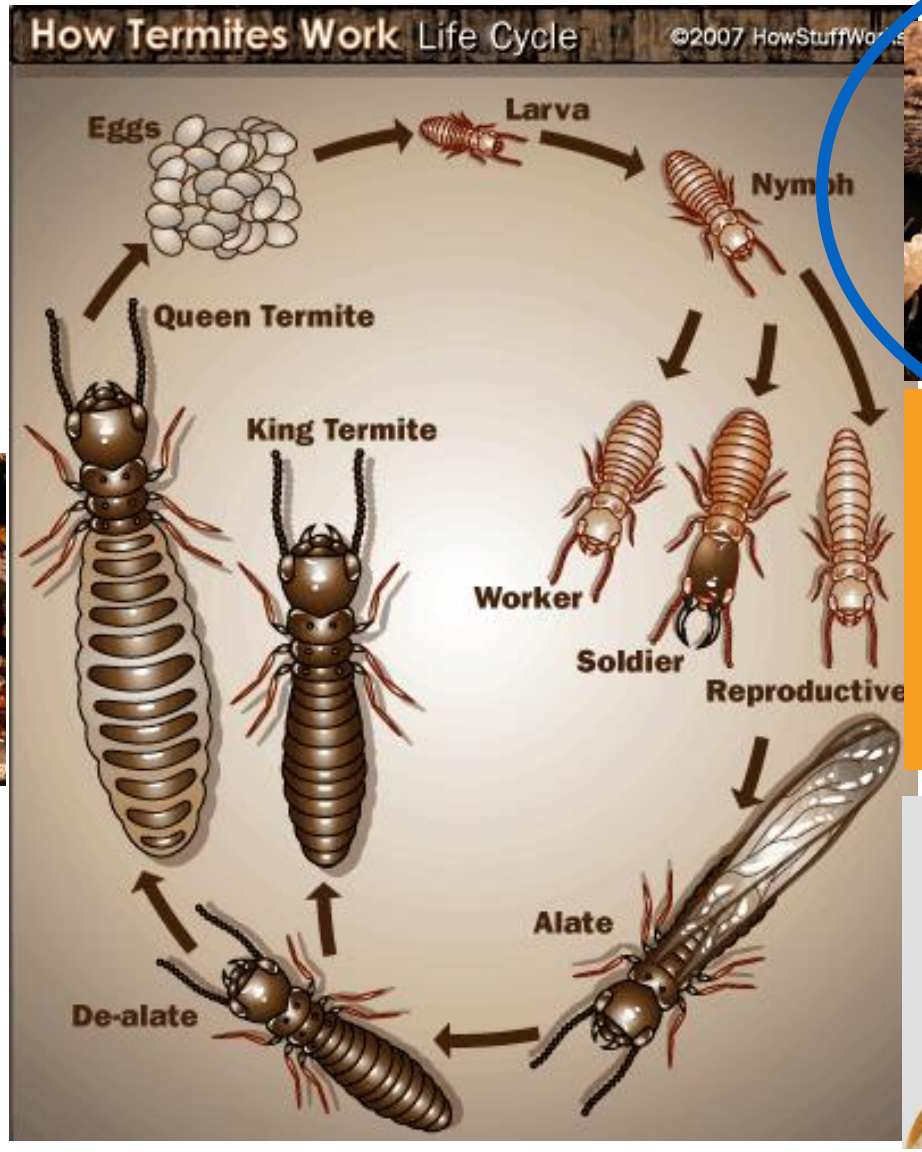
Hemicellulose degrading enzymes



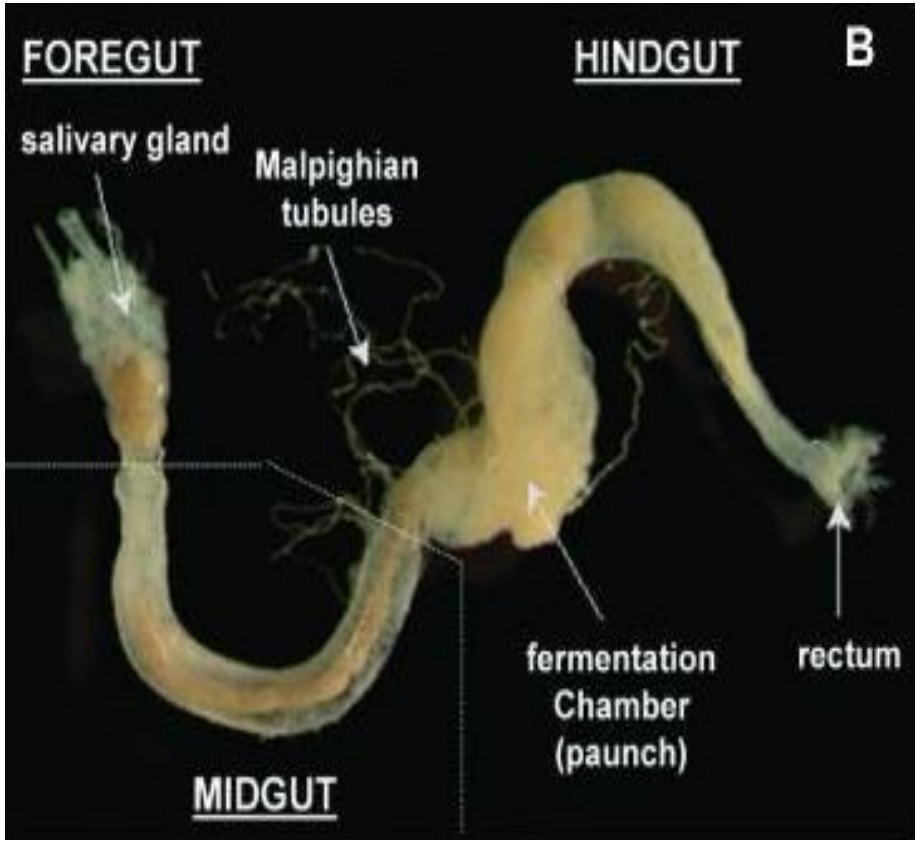
Research subjects

Termite

Termite life cycle



Structure of termite gut



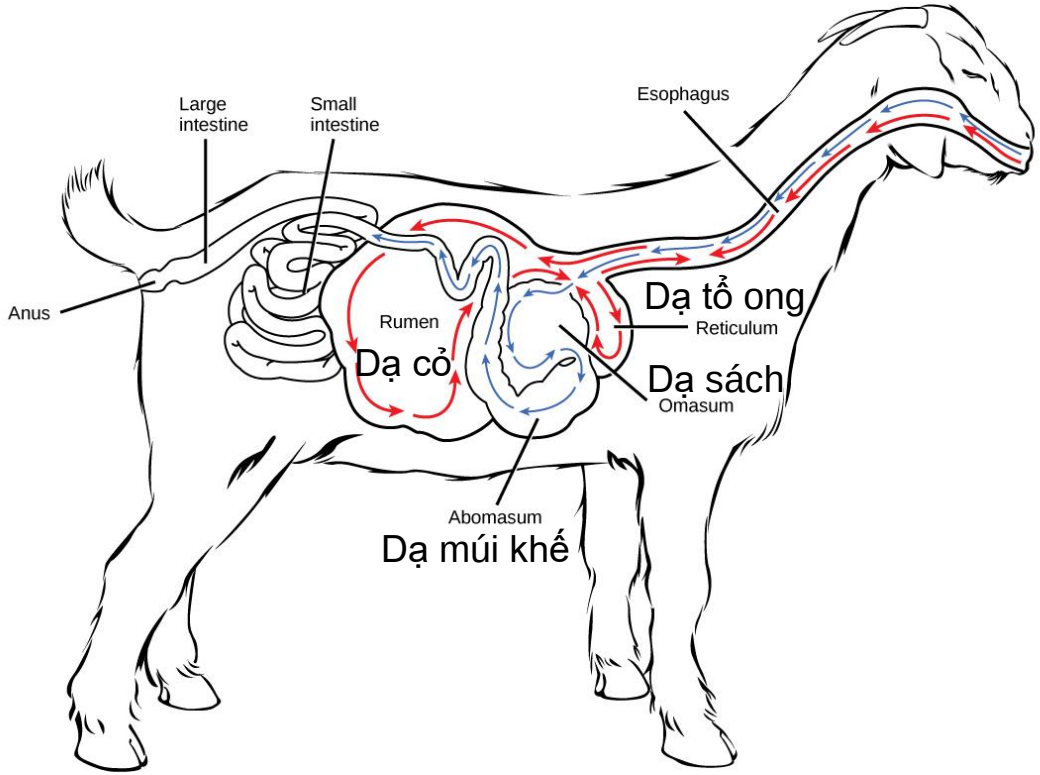
***Hình**: Cấu tạo ruột mối: ruột trước (foregut); ruột giữa (midgut) và ruột sau (hindgut)* []

Goat

Native goat (Co)

Bach Thao

10 samples in Ninh Binh, Thanh Hoa



Goat: Biomass mini factory.

Complex and diverse microbial community:

- Breakdown fibers
- Make short fatty acid chains
- Provide energy

White-rot fungi (nấm mục trắng)

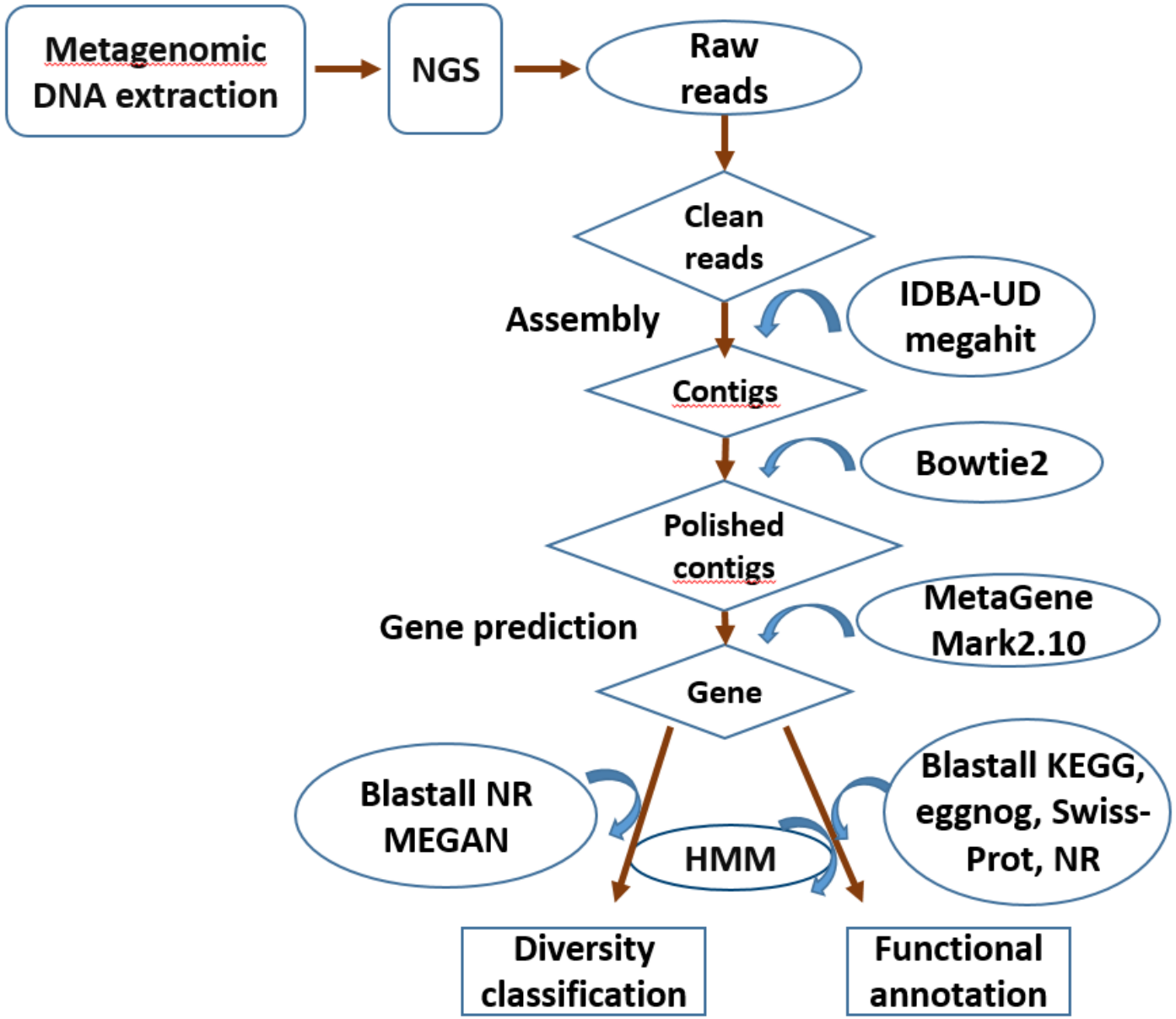
(Humus: Microbiome surrounding white-rot fungi)

Figure 1. Some pictures at the humus sample collection points in Cuc Phuong tropical forest.

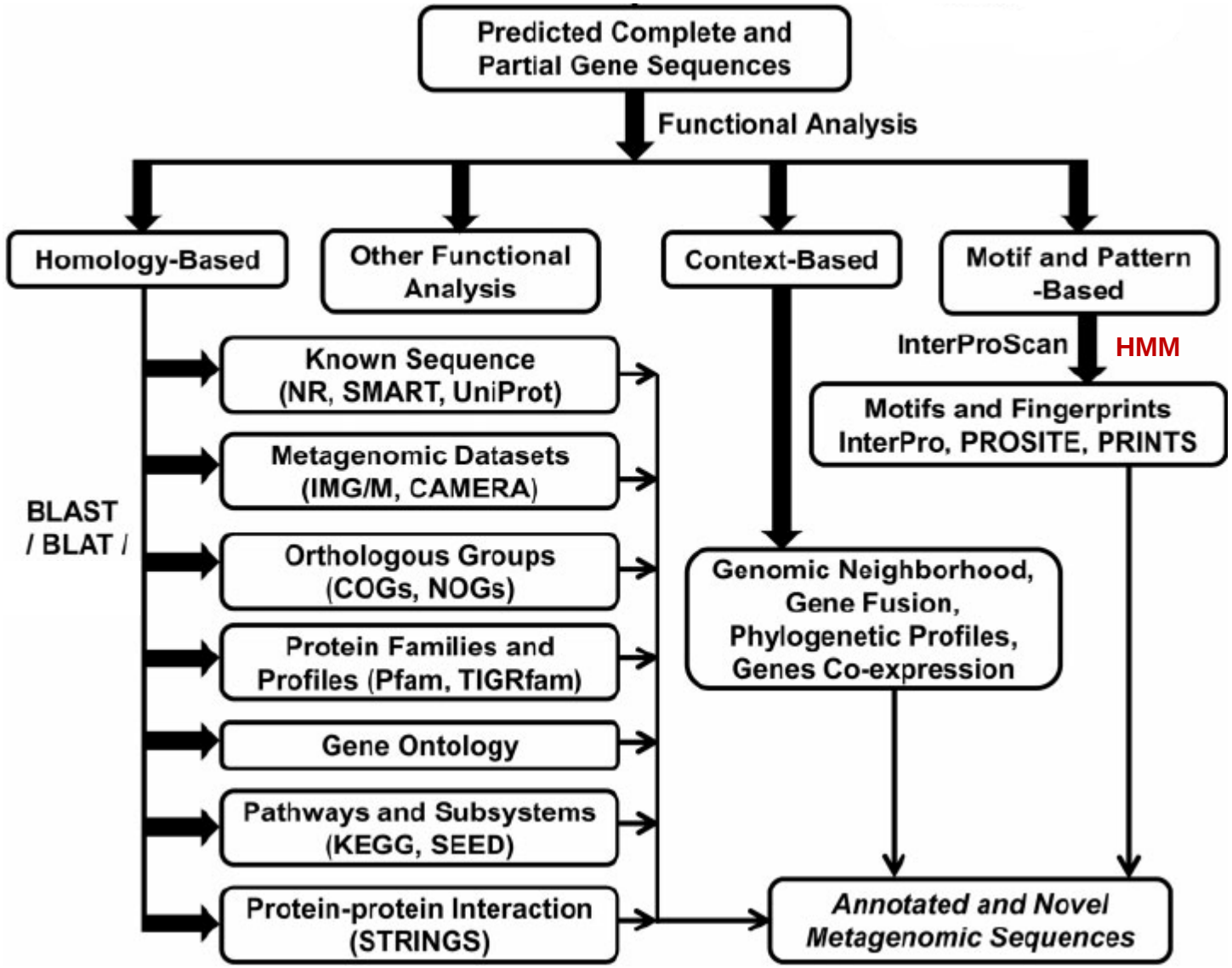
45 samples were harvested from the National rain forest Cuc Phuong, Ninh Binh

2. Metagenomic DNA data mining

Pipeline of metagenomic DNA data mining



Functional assignment of metagenomic data

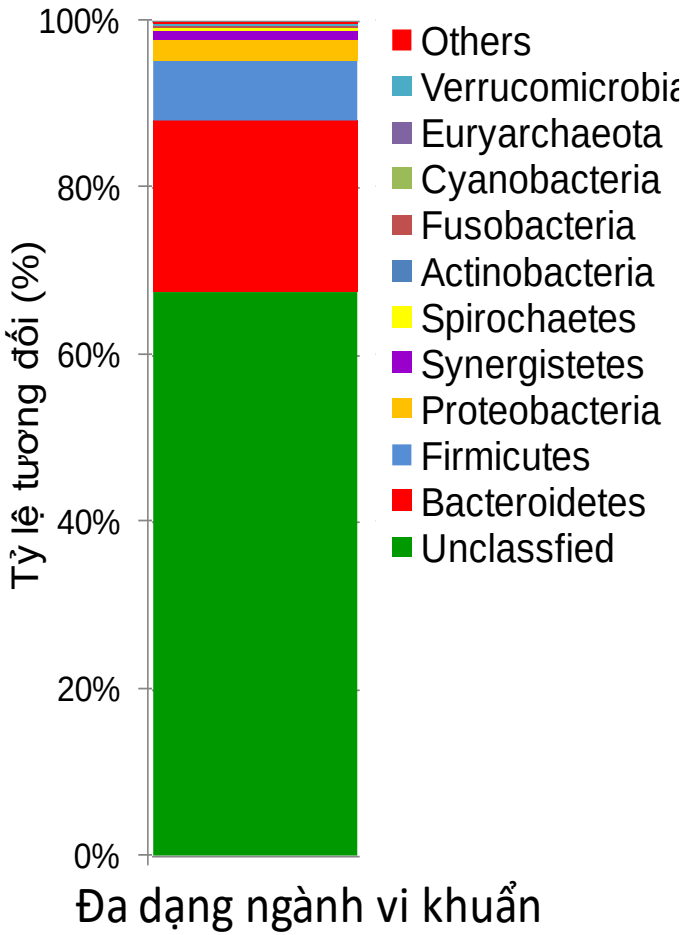


Prakash, Tulika & Taylor, Todd. (2012). Functional assignment of metagenomic data: Challenges and applications. Briefings in bioinformatics. 13. 10.1093/bib/bbs033.

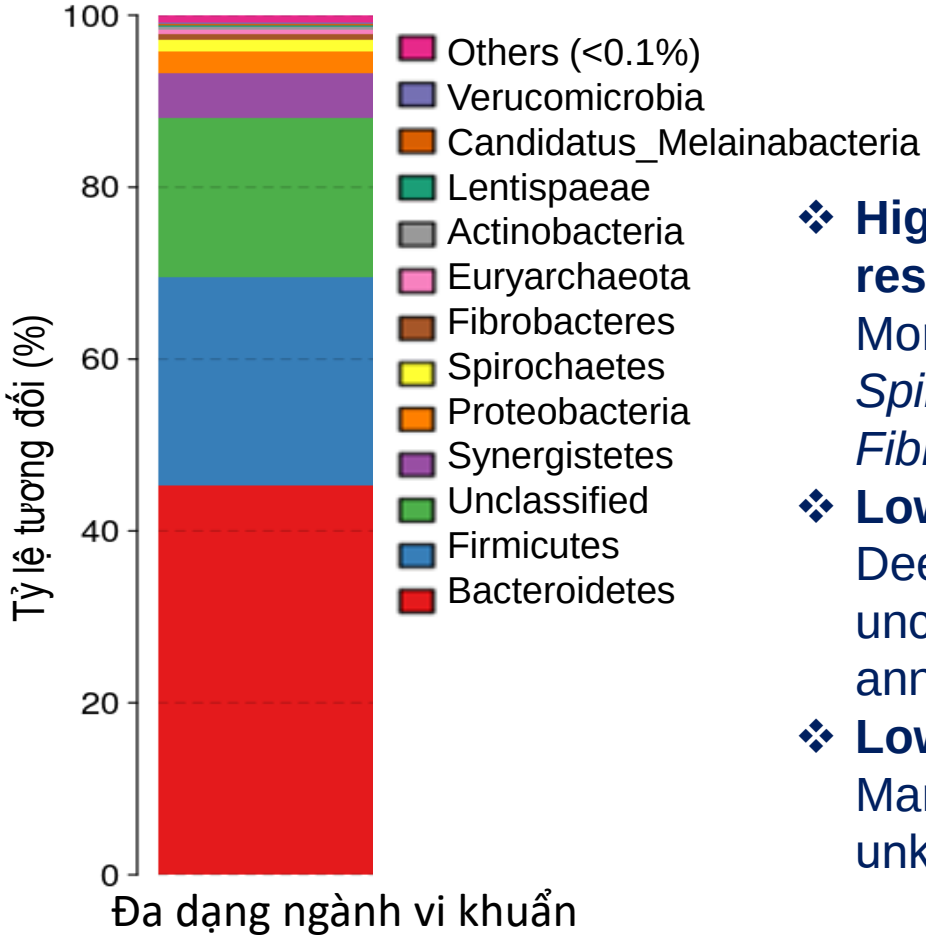
Summary of metagenomic data of termite's gut, goat's rumen and humus samples

Category	Termite (~ 5,4 Gb)	Goat (~ 50 Gb)	Humus (~ 50 Gb)
Total reads	54,316,028	324,080,208	345,471,086
Quality of sequencing Q30 (%)		94,59	94,49
Number of contigs	79,262	3,411,867	2,611,883
Mapping rate (%)	40.09	64.22	62.81
Total length of ORFs(Kbp)	78,271	2,828,583	2,074,265
Number of ORFs	**125,431**	**5,367,270**	**4,104,872**
Annotated genes (%) by Nr	96.858 (77.22%)	4,360,747 (81.25%)	3,923,046 (95.57%)
Annotated genes (%) by SwissProt		2.169.982 (40.43%)	2.382.630 (58.04%)
Annotated genes (%) by KEGG	74.795 (59.63%)	2.700.371 (50.31%)	2.809.791 (68.45%)
Annotated genes (%) by eggNOG	85,444 (68.12%)	3,414,550 (63.62%)	3,279,853 (79.90%)

Why sequencing depth matters?



Depth 8,6 Gbp



Depth 48,6 Gbp

- **Higher depth = better microbial resolution**
  More phyla detected (e.g., *Spirochaetes*, *Synergistetes*, *Fibrobacteres*).
- **Lower unclassified proportion**
  Deeper sequencing reduces unclassified reads, enabling better annotation
- **Low depth = wasted data**
  Many sequences remain functionally unknown → inefficient use of DNA data

Low sequencing depth → high proportion of unclassified reads → loss of functional information → waste of DNA resources.

Diversity of microbial groups in termite's gut, goat's rumen and humus samples

		ORFs	(%)	Phylum (Ngành)	Class (Lớp)	Order (Bộ)	Family (Họ)	Genus (Chi)	Species (Loài)	
Termite (~5,4 Gb)	Bacteria	125.431	98,52	8	41	97	217	628	1368	218%
	Archaea	658	0,52	4	10	15	21	47	61	
	Eukaryota	909	0,71	20	14	25	33	43	24	
	Viruses	313	0,25	0	0	1	11	13	7	
	Total	127311	100	32	65	138	282	731	1460	
Goat (~50 Gb)	Bacteria	5,283,997	98.45	13	85	182	430	2217	946	42%
	Archaea	58,196	1.35	15	26	23	37	117	39	
	Eukaryota	19,264	0.45	11	38	85	191	349	352	
	Viruses	5,813	0.13	0	0	4	29	143	81	
	Total	5,367,270	100	39	149	294	687	2826	1418	
Humus (~50 Gb)	Bacteria	3,884,879	99.69	7	83	170	406	1971	738	37%
	Archaea	293	0.01	9	12	18	23	50	8	
	Eukaryota	1,144	0.03	7	26	46	79	113	86	
	Viruses	10,565	0.27	0	0	2	14	101	84	
	Total	4,104,872	100	23	121	226	232	2235	916	

Summary of bacterial phylum abundance in termite's gut, goat's rumen and humus

Phylum	Termite (%)	Goat (%)	Humus (%)
Firmicutes	22,48% (1)	35,99 (1)	1,40 (4)
Proteobacteria	17,84% (2)	3,04 (4)	75,68 (1)
Spirochaetes	17,40% (3)	2,07	
Bacteroidetes	11,60% (4)	23,35 (2)	13,11 (2)
Synergistetes	4,27%	3,27 (3)	
Planctomycetes	1,14%		0,10
Actinobacteria	0,48%	0,45	1,60 (3)
Euryachaeota	0,40	1,07	
Lentispaerae		0,53	
Acidobacteria			0,80
Verrucomicrobia			0,32
...			

KEGG pathway annotation of humus sample genes

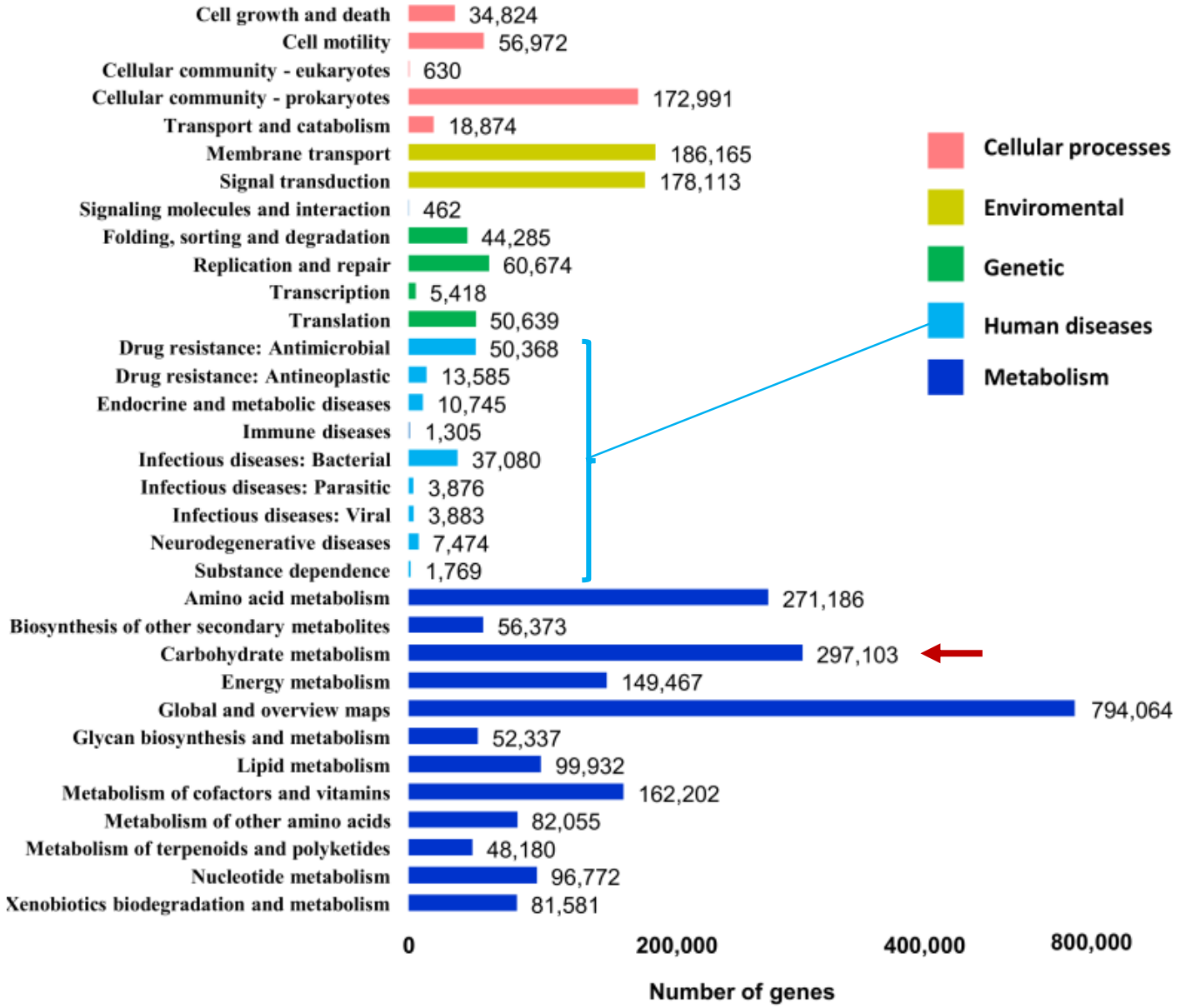
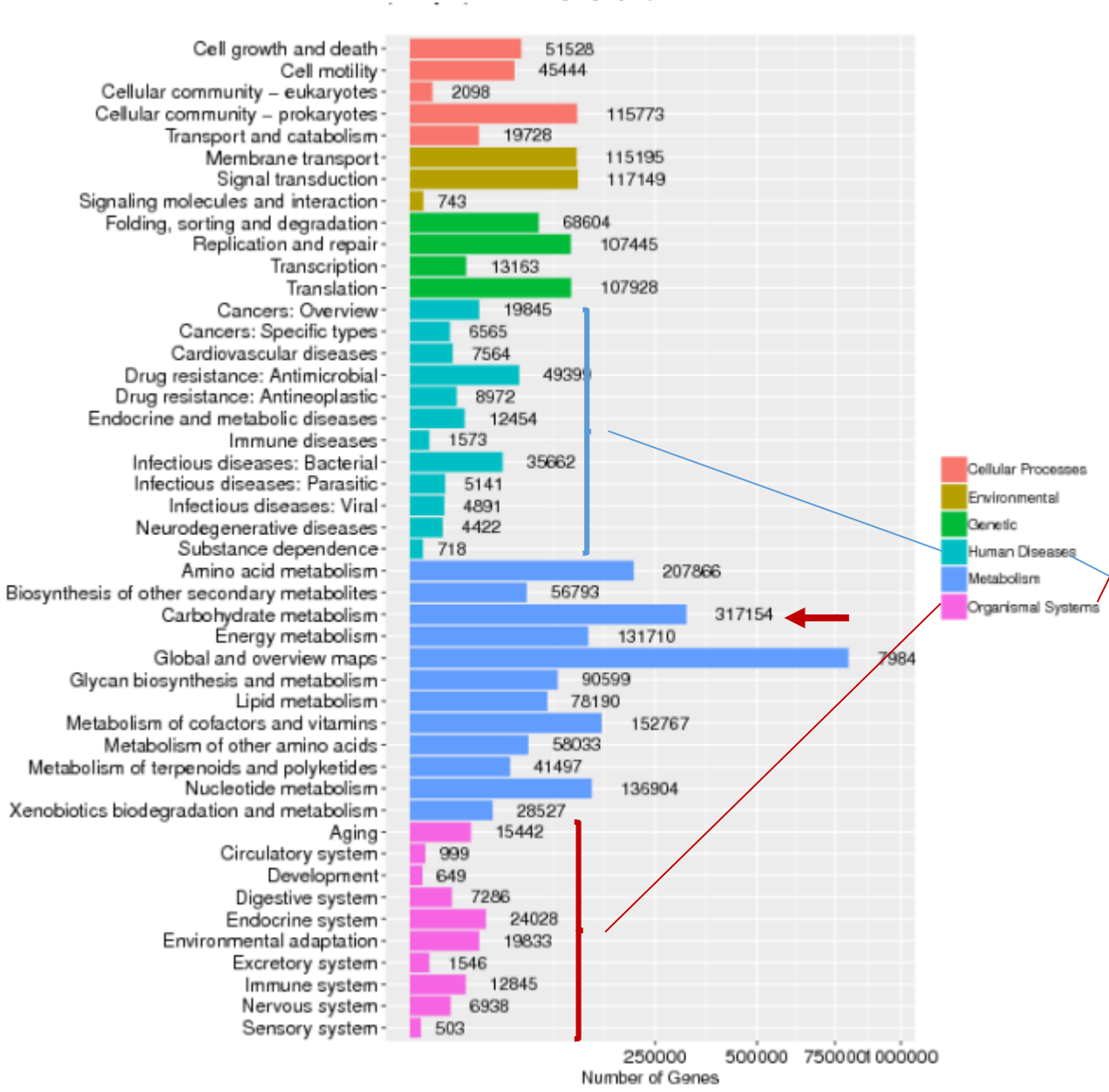


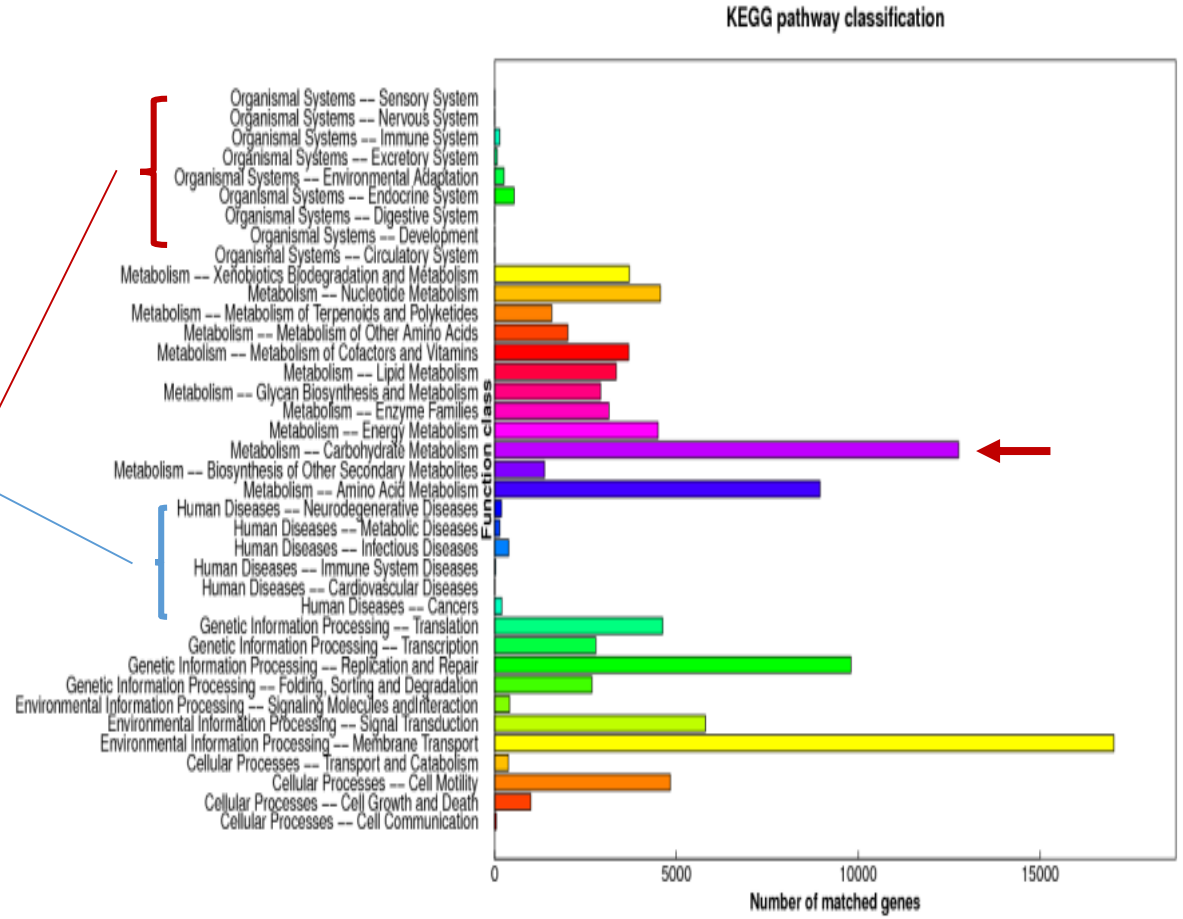
Figure 4. Summary of KEGG annotation. X-axis represents the number of genes that annotated each pathway, and y-axis lists annotated pathways in the particular subclass.

KEGG pathway annotation of termite's gut and goat's rumen genes

Goat



Termite



Mining metagenomic DNA databases for lignocellulose degradation enzymes by sequence similarity alignment/sequence based search (SSA/SBS)

Samples	Number of ORFs	Pretrea-tment (ORFs)	Cellulase (ORFs)	Hemi-cellulase (ORFs)
Humus	4,104,872	903	8,301	13,026
Goat's rumen	5,367,270	2,670	21,029	43,841

Distribution of cellulases, hemicellulases and pretreatment enzymes of bacterial groups in goat's rumen

Phylum

Cellulase	17,566	100%	Hemicellulase	39,827	100%	Pretreatment	2,430	100%
Firmicutes	9152	52.10	Bacteroidetes	23448	58.87	Bacteroidetes	1651	67.94
Bacteroidetes	6849	38.99	Firmicutes	13210	33.17	Firmicutes	582	23.95
Fibrobacteres	403	2.29	Spirochaetes	823	2.07	Spirochaetes	51	2.10
Spirochaetes	351	2.00	Lentisphaerae	593	1.49	Proteobacteria	40	1.65
Proteobacteria	246	1.40	Fibrobacteres	466	1.17	Fibrobacteres	27	1.11
Actinobacteria	199	1.13	Verrucomicrobia	356	0.89	Actinobacteria	13	0.53
Lentisphaerae	176	1.00	Actinobacteria	270	0.68	Lentisphaerae	11	0.45
Verrucomicrobia	59	0.34	Proteobacteria	239	0.60	Candidatus Aminicenantes	9	0.37

Genus

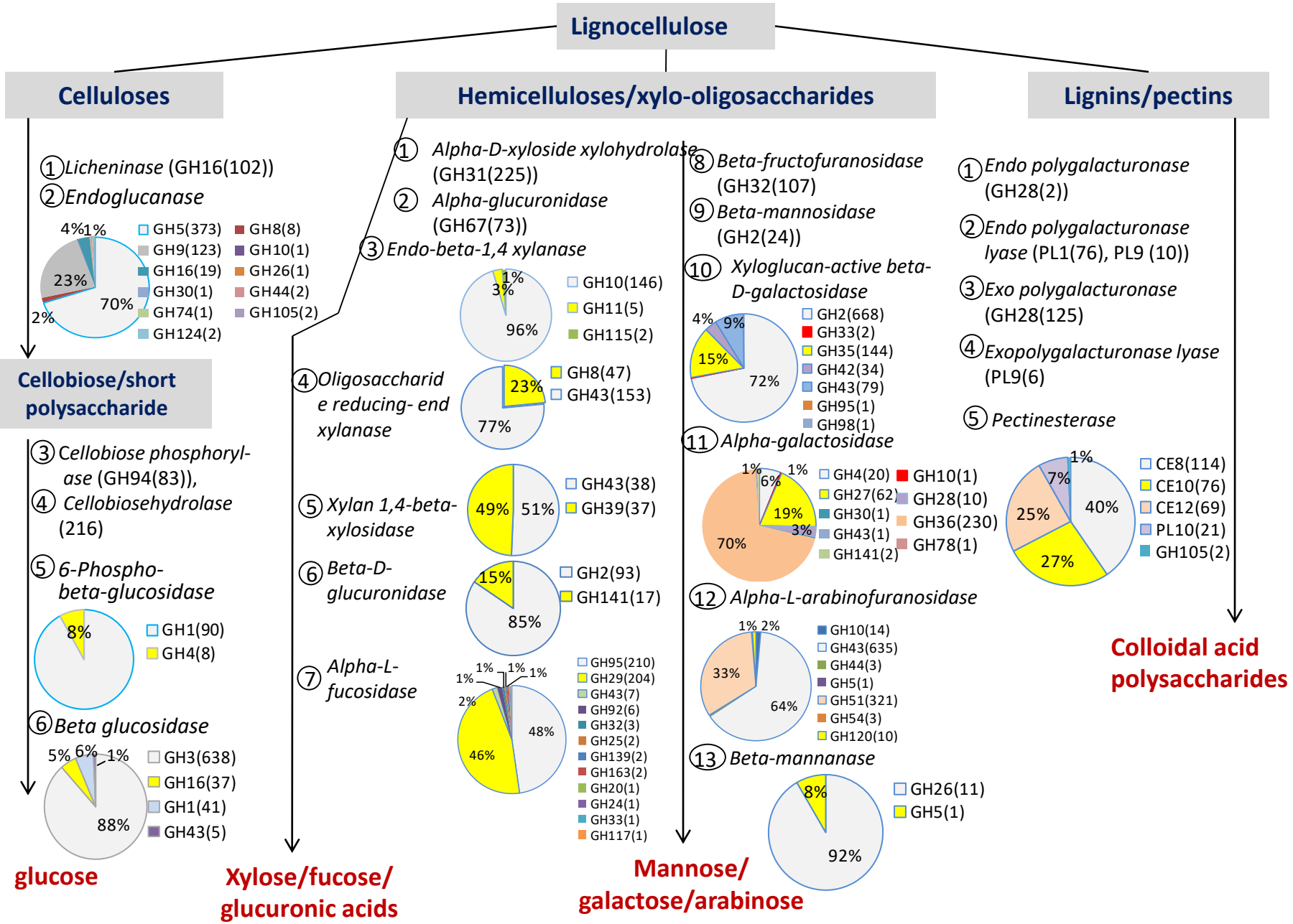
Cellulase	10,659	100%	Hemicellulase	24,644	100%	Pretreatment	1,548	100%
Prevotella	**3603**	**33.80**	***Prevotella***	**13135**	**53.30**	***Prevotella***	**812**	**52.45**
Ruminococcus	1564	14.67	*Bacteroides*	1851	7.51	*Bacteroides*	104	6.72
Butyrivibrio	978	9.18	*Butyrivibrio*	1759	7.14	*Ruminococcus*	100	6.46
Bacteroides	576	5.40	*Ruminococcus*	1409	5.72	*Alistipes*	77	4.97
Treponema	338	3.17	*Treponema*	768	3.12	*Butyrivibrio*	66	4.26
Selenomonas	328	3.08	*Alistipes*	689	2.80	*Treponema*	49	3.17
Fibrobacter	319	2.99	*Clostridium*	436	1.77	*Acidaminococcus*	27	1.74
Clostridium	240	2.25	*Selenomonas*	385	1.56	*Fibrobacter*	23	1.49
Faecalibacterium	230	2.16	*Fibrobacter*	365	1.48	*Selenomonas*	22	1.42
Alistipes	186	1.75	*Eubacterium*	264	1.07	*Muribaculum*	20	1.29
Eubacterium	170	1.59	*Paenibacillus*	263	1.07	*Clostridium*	20	1.29
Pseudobutyrivibrio	126	1.18	*Paraprevotella*	238	0.97	*Lewinella*	16	1.03
Lachnoclostridium	121	1.14	*Faecalibacterium*	201	0.82	*Paenibacillus*	15	0.97

Mining lignocellulases from metagenomic DNA database of goat's rumen

Pretreatment & others	EC ID	4769
Expansin		33
Feruloyl esterase	3.1.1.73	92
Laccase	1.10.3.2	9
Lignin peroxidase	1.11.1.14	0
Lytic polysaccharide monooxygenase	1.14.99.56	11
Manganese peroxidase	1.11.1.13	0
Pectinesterases	3.1.1.11	2525
Exopolygalacturonase	3.2.1.67	635
Exopolygalacturonase lyase	4.2.2.9	59
Endopolygalacturonase	3.2.1.15	27
Cellulases	**EC**	**21029**
Licheninase	3.2.1.73	281
Endoglucanase	3.2.1.4	7368
β-glucosidase	3.2.1.21	10444
6-phospho-beta-glucosidase	3.2.1.86	1281

Hemicellulases	EC ID	41756
Cellobiohydrolase	3.2.1.91	216
Acetylxylan esterases	3.1.1.72	4
α-D-Xyloside xylohydrolase	3.2.1.177	2833
α-Glucuronidase	3.2.1.139	561
β-D-Glucuronidase	3.2.1.31	888
β-Fructofuranosidase	3.2.1.26	1106
β-Mannosidase	3.2.1.25	659
Endo-transglycosylase/hydrolases	2.4.1.207	2
Endo-β-1,4 xylanases	3.2.1.8	3400
Oligosaccharide reducing-end xylanase	3.2.1.156	1213
Xylan 1,4-beta-xylosidase	3.2.1.37	1018
Xyloglucan-active β-D-galactosidase	3.2.1.23	11690
α-Galactosidase	3.2.1.22	4020
α-L-Arabinofuranosidases	3.2.1.55	6229
α-L-Fucosidase	3.2.1.51	6896
β-Mannanase	**3.2.1.78**	**1237**
Cellobiose phosphorylase	2.4.1.20	1439

Overview picture of GHs related to lignocellulose degradation of bacteria in goat's rumen



The microbial community in the goat rumen possesses a **rich and specialized arsenal of lignocellulose-degrading enzymes**, including:

- **Cellulose** is efficiently broken down by GH5, GH9, and GH3.
- **Hemicellulose** is targeted by a much more diverse network of enzymes, reflecting evolutionary adaptation to maximize carbon extraction from complex structures.
- **Lignin/pectin** enzymes are less abundant but still present, indicating a certain level of degradation capability.

➞ The rumen microbiome plays a vital role in supporting the ruminant's ability to utilize fibrous plant materials.

Biological significance of GH diversity:

- The presence of multiple GHs with similar functions is an **evolutionary strategy** enabling microbes to adapt to various substrates.
- Some bacteria may encode **redundant GHs**, which enhances degradation efficiency and ensures functional robustness in the complex rumen environment.

Distribution of pretreatment enzymes, cellulases and hemicellulases in bacterial groups in humus surrounding white spot fungi

Description	Num
Pretreatment	907
Cellulase	8301
Hemicellulase	13018

Pretreatment

Phylum	Order	Num
Bacteroidetes	Flavobacteriales	271
Bacteroidetes	Sphingobacteriales	178
Proteobacteria	Enterobacterales	158
Proteobacteria	Xanthomonadales	64
Proteobacteria	Pseudomonadales	34
Proteobacteria	Caulobacterales	28
Bacteroidetes	Bacteroidales	21
Proteobacteria	Burkholderiales	21
Bacteroidetes	Cytophagales	14

Cellulase

Phylum	Order	Num
Proteobacteria	Enterobacterales	2316
Bacteroidetes	Flavobacteriales	915
Proteobacteria	Xanthomonadales	727
Proteobacteria	Sphingomonadales	603
Bacteroidetes	Sphingobacteriales	572
Proteobacteria	Pseudomonadales	335
Proteobacteria	Rhizobiales	271
Proteobacteria	Burkholderiales	267
Actinobacteria	Micrococcales	256
Bacteroidetes	Chitinophagales	233
Proteobacteria	Caulobacterales	182
Firmicutes	Bacillales	144
Bacteroidetes	Cytophagales	141
Bacteroidetes	Bacteroidales	130
Firmicutes	Lactobacillales	101

Hemicellulase

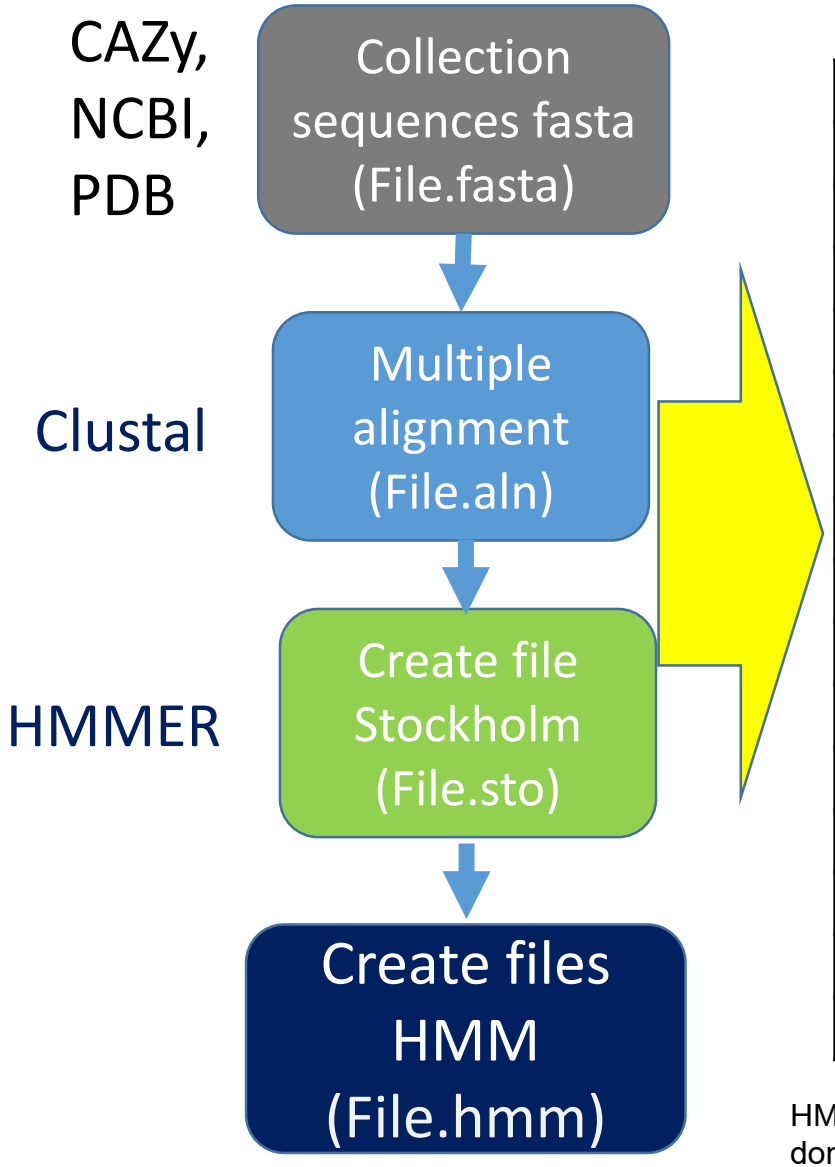
Phylum	Order	Num
Bacteroidetes	Flavobacteriales	2176
Proteobacteria	Enterobacterales	1986
Bacteroidetes	Sphingobacteriales	1907
Proteobacteria	Sphingomonadales	1121
Proteobacteria	Xanthomonadales	908
Bacteroidetes	Chitinophagales	661
Proteobacteria	Rhizobiales	557
Bacteroidetes	Bacteroidales	455
Proteobacteria	Burkholderiales	424
Proteobacteria	Caulobacterales	303
Firmicutes	Bacillales	298
Acidobacteria	Acidobacteriales	287
Bacteroidetes	Cytophagales	231
Actinobacteria	Micrococcales	229
Firmicutes	Lactobacillales	110

Mining lignocellulases from metagenomic DNA database of humus surrounding white spot fungi

Pretreatment & others	EC ID	4769
Expansin		7
Feruloyl esterase	3.1.1.73	75
Laccase	1.10.3.2	10
Lignin peroxidase	1.11.1.14	0
Lytic polysaccharide monooxygenase	1.14.99.56	0
Manganese peroxidase	1.11.1.13	0
Pectinesterase	3.1.1.11	815
Endopolygalacturonase	3.2.1.15	38
Endopolygalacturonase lyase	4.2.2.2	341
Exopolygalacturonase	3.2.1.67	223
Cellulases	**EC**	**21029**
β-Glucosidase	3.2.1.21	4272
Endoglucanase	3.2.1.4	2216
6-Phospho-beta-glucosidase	3.2.1.86	1718
Cellobiohydrolase	3.2.1.91	73

Hemicellulases	EC ID	41756
Acetyl mannan esterase	3.1.1.6	0
Acetylxylan esterase	3.1.1.72	1
α-D-Xyloside xylohydrolase	3.2.1.177	762
α-Glucuronidase	3.2.1.139	161
β-D-Glucuronidase	3.2.1.31	227
β-Fructofuranosidase	3.2.1.26	255
β-Mannosidase	3.2.1.25	611
Endo-transglycosylase/hydrolase	2.4.1.207	2
Endo-β-1,4 xylanase	3.2.1.8	885
Oligosaccharide reducing-end xylanase	3.2.1.156	552
Xylan 1,4-beta-xylosidase	3.2.1.37	659
Xyloglucan-active β-D-galactosidase	3.2.1.23	3288
α-Galactosidase	3.2.1.22	1033
α-L-Arabinofuranosidase	3.2.1.55	1016
α-L-Fucosidase	3.2.1.51	2279
β-Mannanase	**3.2.1.78**	**368**
Cellobiose phosphorylase	2.4.1.20	22
Licheninase	3.2.1.73	175

Creating profile HMM for mining lignocellulose degradation genes from metagenome DNA databases



NN	Candidate genes	NN	Candidate genes
1	Lignin peroxidase (LiP)	16	Acetyl xylan esterase
2	Hydrogen peroxide oxidoreductase	17	Feruloyl esterase
3	Manganese peroxidase (MnP)	18	Xyloglucanse
4	Versatile peroxidase (VP)	19	Mannanase
5	Laccase	20	β-mannosidase
6	Expansin	21	Arabinase
7	Lytic polysaccharide monooxygenase	22	Galactanase
8	Cellobiohydrolase	23	Polygalacturonase
9	Endo-1,4-glucanase	24	β-glucuronidase
10	β-glucosidase	25	Glucuronyl esterase
11	Endo-xylanase	26	CBM(1-84)
12	β-xylosidase	27	Fibronectin 3-like domain (FN3)
13	α-L-arabinofuranosidase	28	Dockerin
14	α-glucuronidase	29	Immunoglobulin-like domain (Ig)
15	Licheninase		

HMMER is software package that provides tools for making probabilistic models of protein and DNA sequence domain families called profile hidden Markov models for using to annotate new sequences…

Genes coding for lignocellulose degradation enzymes from metagenomic DNA databases mined by profile HMM

	Goat	Humus
Cellulase	**5236**	**1928**
Endoglucanase	2009	557
Cellobiohydrolase	1645	253
β-Glucosidase	1582	1118
Hemicellulase	**19597**	**4842**
α-L-arabinofuranosidase	2646	431
α Glucurinindase (GH76N)	615	102
Arabinanase (GH43)	1894	343
Axetylxylanesterase (AXE1)	219	79
β_Glucuronidase	4171	1044
β_Xylosidase	3276	945
β-Mannosidase (GH2)	1890	594
Feruloyl esterase	7	53
Galactanase	305	17
Glucuronyl esterase	165	22
HPOXRE catalase	1	224

	Goat	Humus
Lichenase	350	290
Mannanase	263	40
Polygalacturonase	211	45
Xylanase (GH44)	**3534**	**599**
Xyloglucanase	14	14
Preatretment & others	15713	6809
MnP, VerP, LiP	0	0
Laccase/MCO	9/350	10/1115
LPMO	1	69
Expansin	36	0
CBM	7285	3163
Dockerin	1734	11
Ig	2688	1178
Fn3	3655	1273

Comparison of mining results of goat metagenomic DNA database by SBS và HMM methods

	SBS	HMM
Cellulase	**18028**	**5236**
Endoglucanase	7368	2009
Cellobiohydrolase	216	1645
β-glucosidase	10444	1582
Hemicellulase	**14723**	**19048**
α-L-arabinofuranosidase	6229	2646
α glucurinindase (GH76N)	561	102
Arabinanase (GH43)	0	1894
Axetylxylanesterase (AXE1)	4	219
β-glucuronidase	888	4171
β-xylosidase	1018	3276
β-mannosidase (GH2)	659	1890
Feruloyl esterase	92	7

	SBS	HMM
Galactanase	0	305
Glucuronyl esterase	0	165
HPOXRE catalase)	0	1
Lichenase	0	350
Mannanase	1237	263
Polygalacturonase	635	211
Xylanase (GH44)	**3400**	**3534**
Xyloglucanase	0	14
Pretreatment & others	**53**	**15713**
MnP, VerP, LiP	0	0
Laccase/MCO	9/0	9/341
LPMO	11	1
Expansin	33	36
CBM	0	7285
Dockerin	0	1734
Ig	0	2688
fn3	0	3655

Comparing mining results of humus metagenome DNA database by SBS và HMM methods

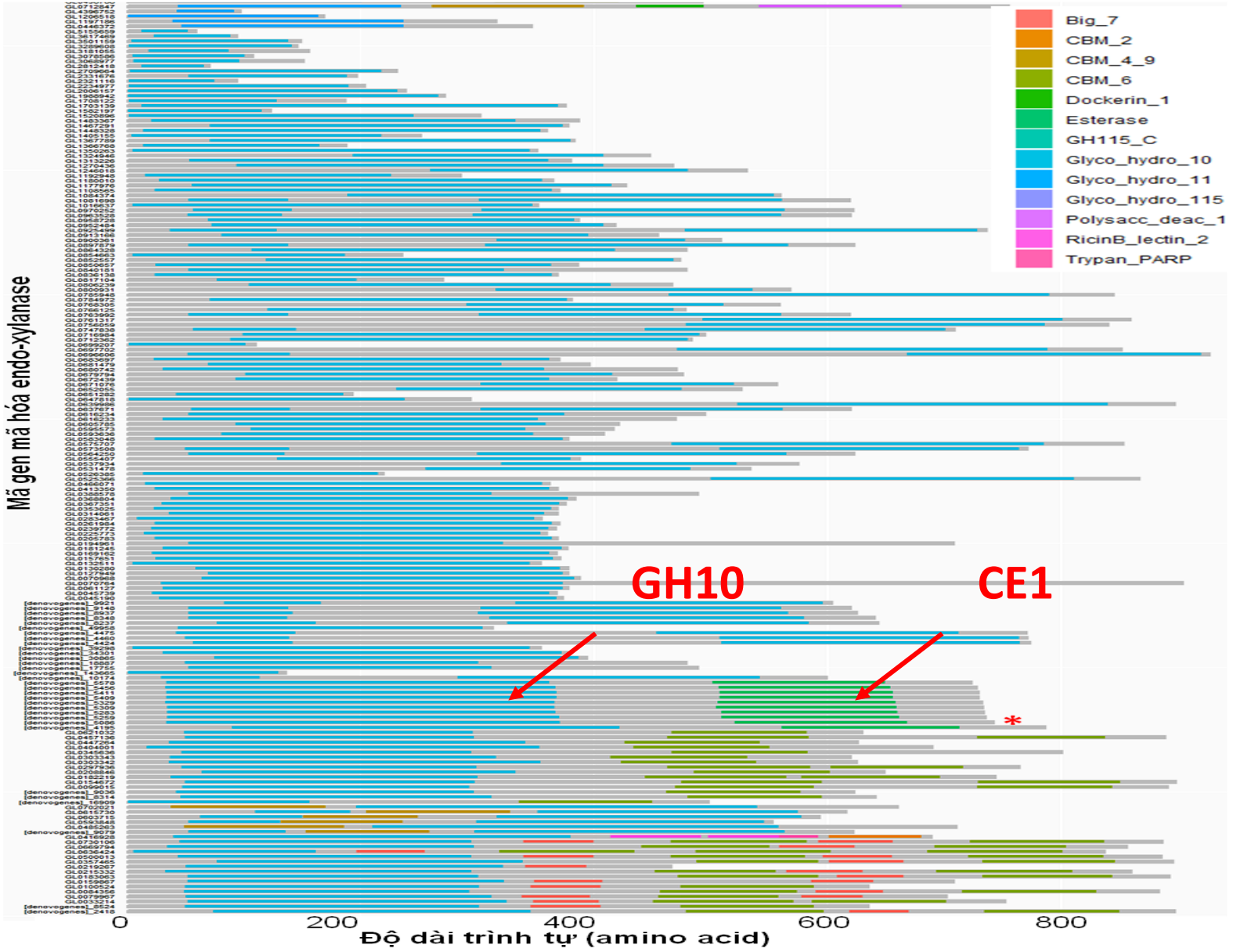
	SBS	HMM
Cellulase	**6488**	**1928**
Endoglucanase	2216	557
Cellobiohydrolase	0	253
β-Glucosidase	4272	1118
Hemicellulase	**3158**	**4842**
α-L-Arabinofuranosidase	1016	431
α Glucurinindase (GH76N)	161	102
Arabinanase (GH43)	0	343
Axetylxylanesterase (AXE1)	1	79
β_Glucuronidase	227	1044
β_Xylosidase	659	945
β-Mannosidase (GH2)	611	594
Feruloyl esterase	75	53
Galactanase	0	17
Glucuronyl esterase	0	22

	SBS	HMM
HPOXRE (catalase)	0	224
Lichenase	0	290
Mannanase	368	40
Polygalacturonase	38	45
Xylanase (GH44)	552	599
Xyloglucanase	0	14
Pretreatment & others	**17**	**6809**
MnP, VerP, LiP	0	0/ 10/110
Laccase/MCO	10/0	5
LPMO	**0**	**69**
Expansin	7	0
CBM	0	3163
Dockerin	0	11
Ig	0	1178
Fn3	0	1273

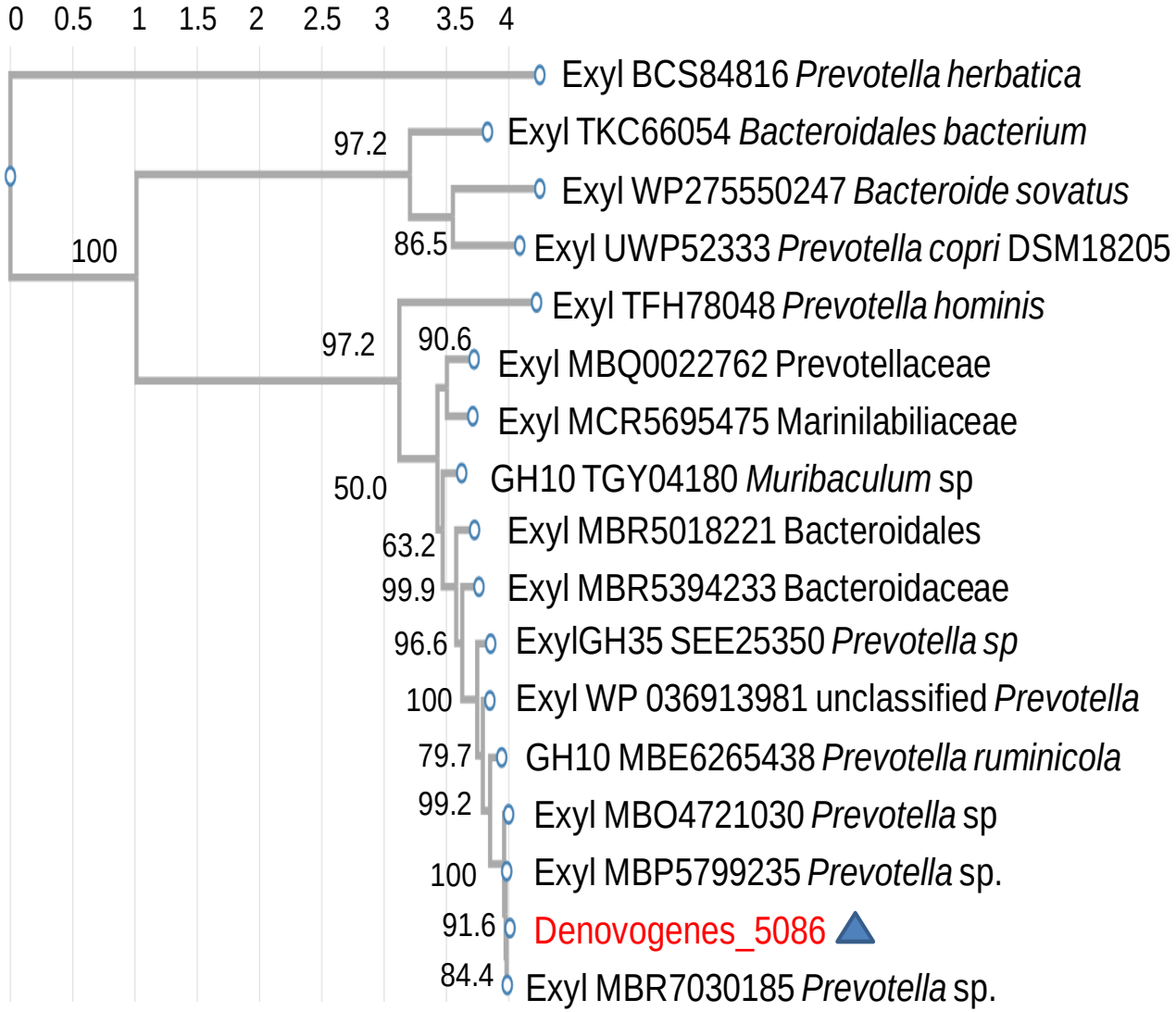
3. Endo-Xylanase

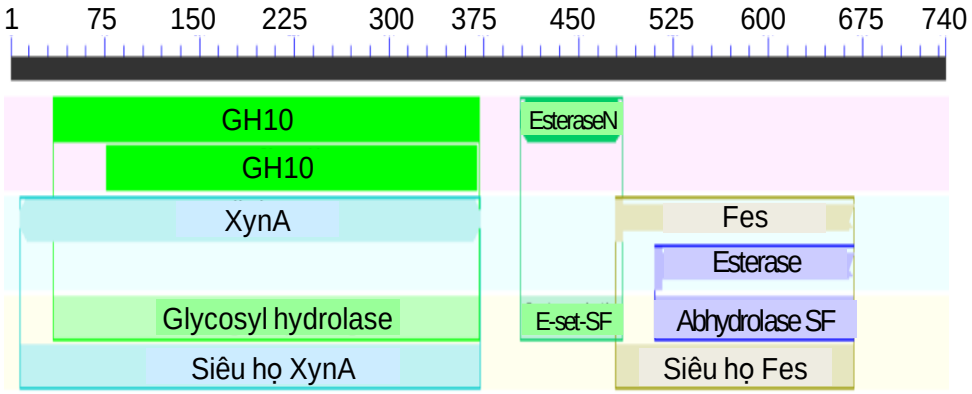
Mining of the gene encoding endo-xylanase

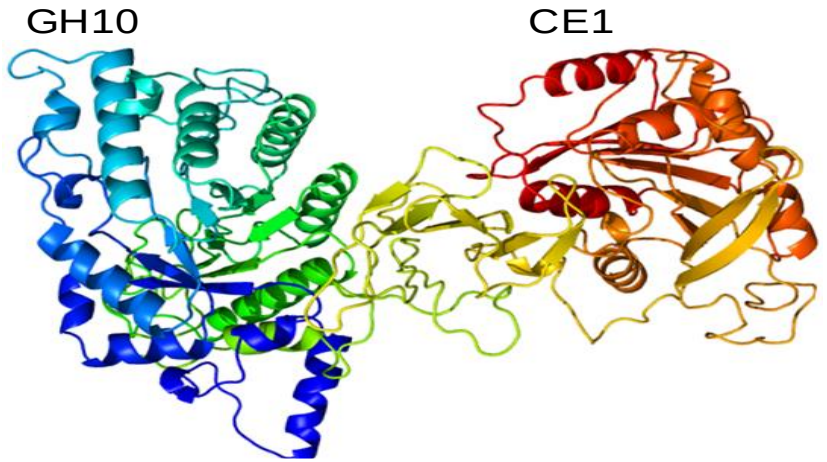
Among the 739 complete genes encoding endo-xylanase, only 185 (~25%) sequences contain a glycosyl hydrolase catalytic domain (InterProscan).



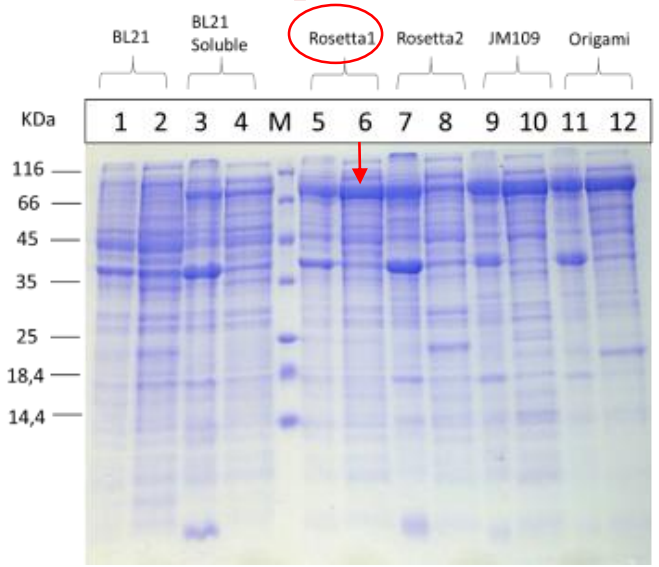
Mining of the gene encoding endo-xylanase



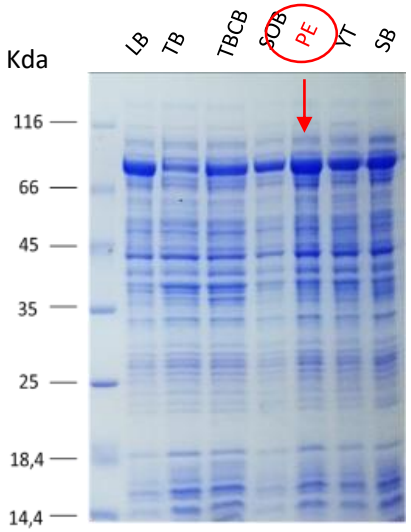




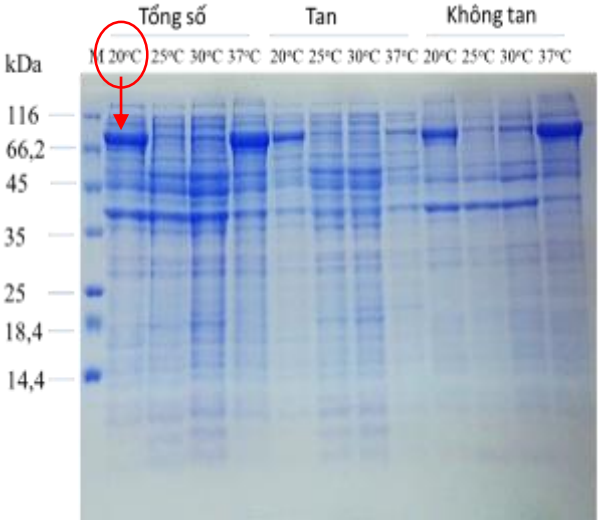
Expression of the endo-xylanase gene



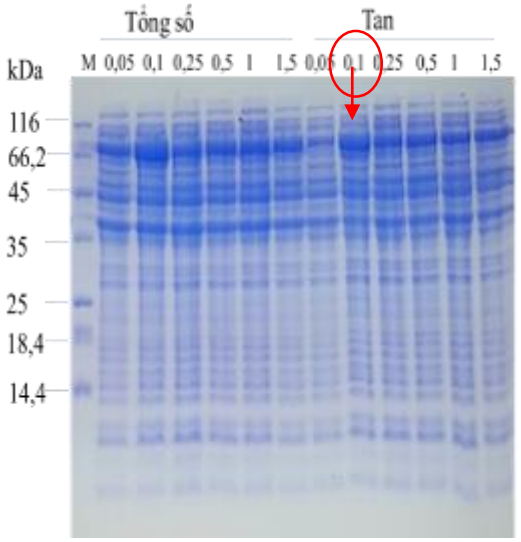
Host cells



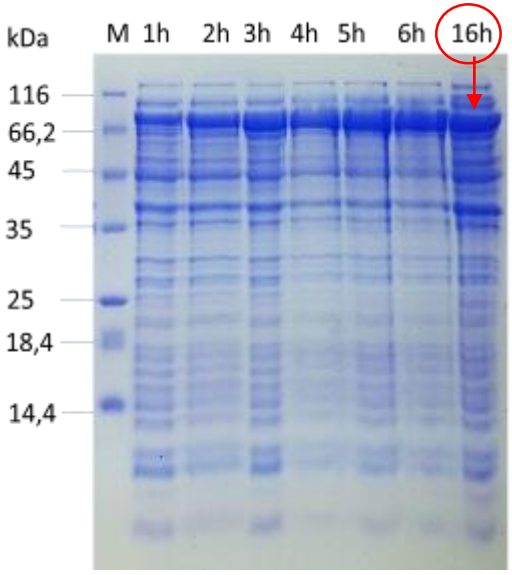
Medium



Temperature



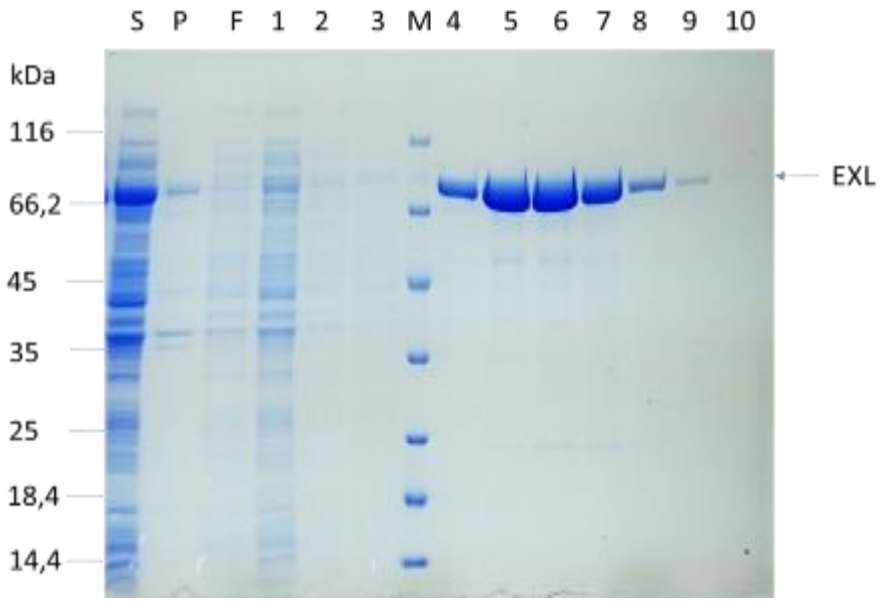
IPTG concentration



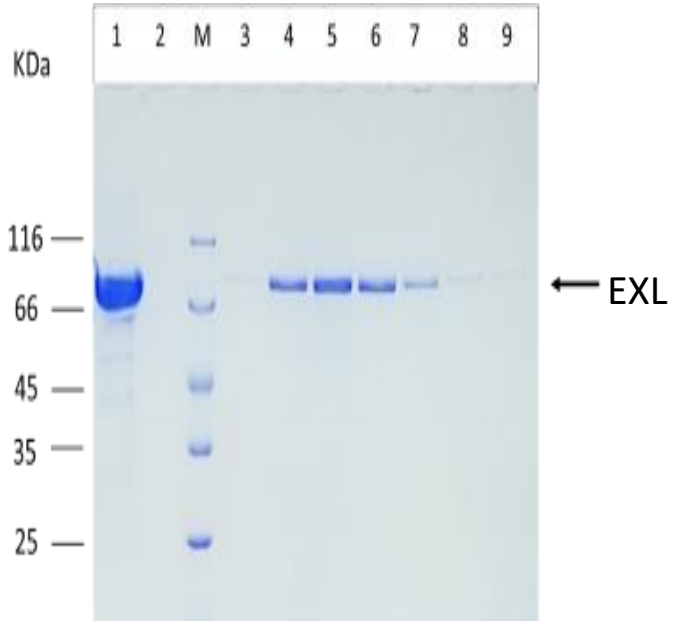
Harvest time

EXL expressed well in Rosetta1, media PE, at 20°C, 0.1 mM IPTG and after induction 16 hours.

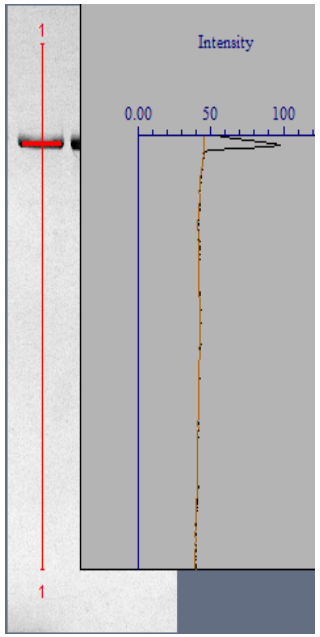
Purification of endo-xylanase



Purification

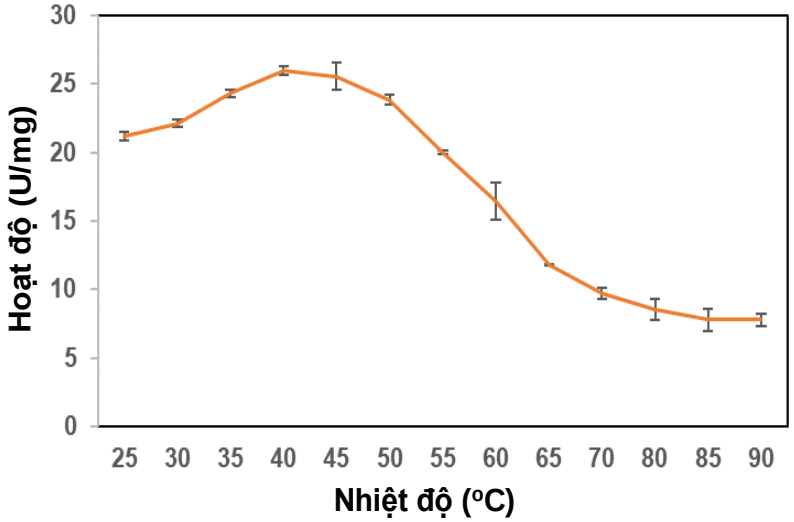


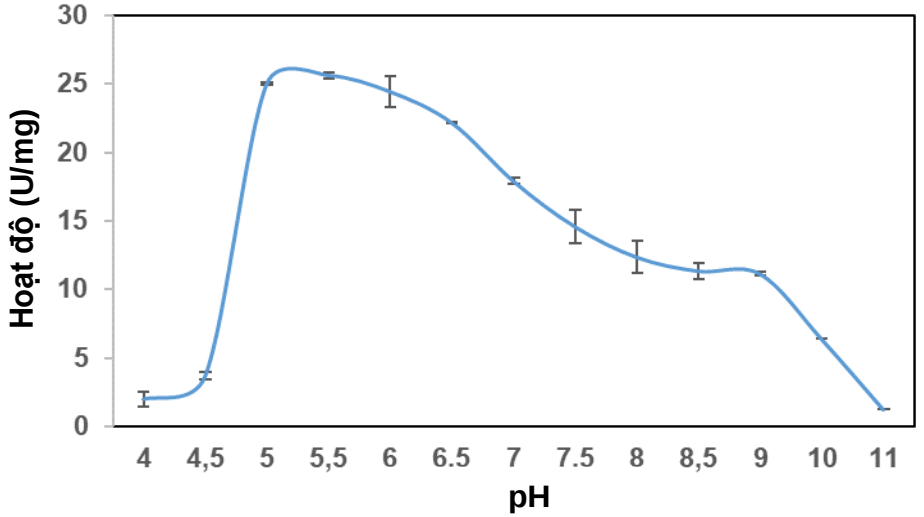
Salt elimination

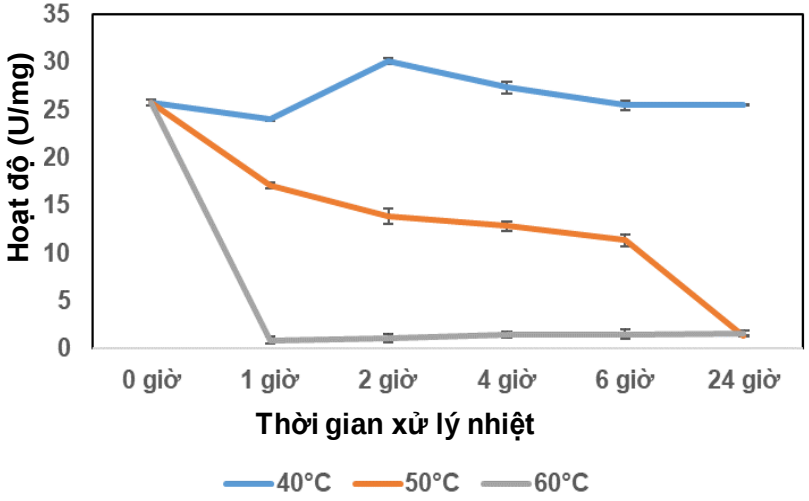


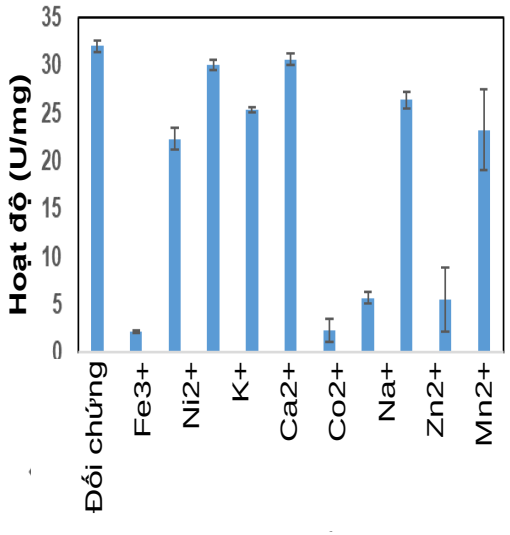
Purity check

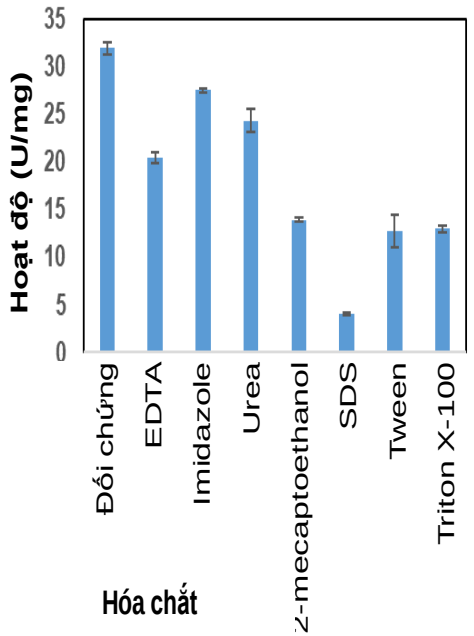
Biochemical properties of endo-xylanase





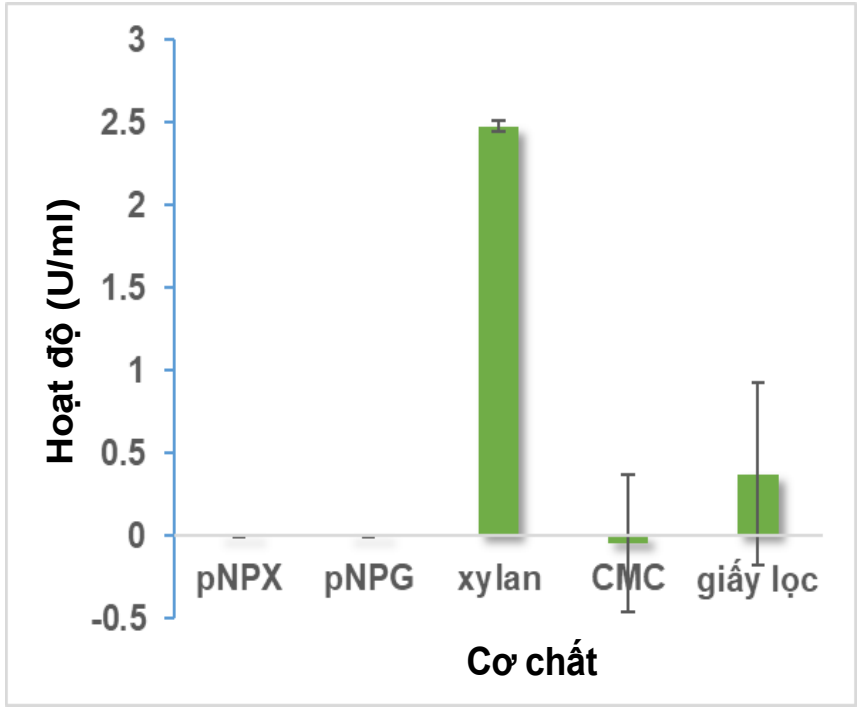


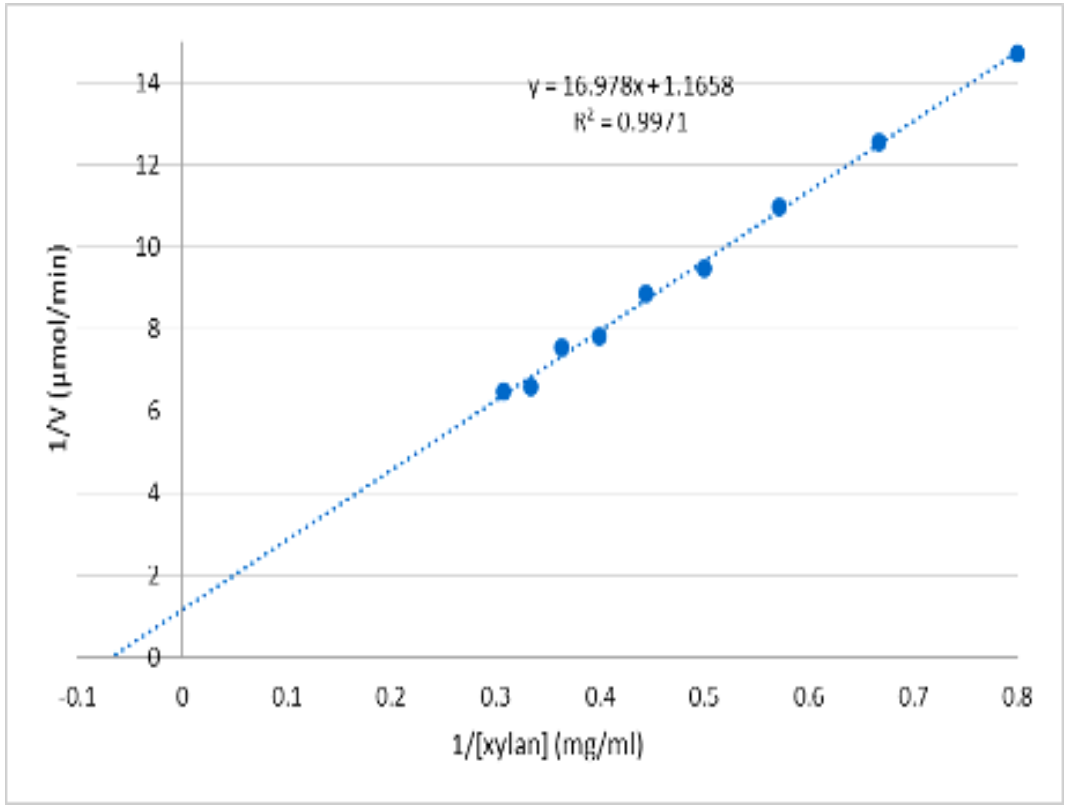




EXL works optimally at 40°C, pH 5.5, is stable at optimal temperature and less stable at higher temperatures, and is less active in environments containing metal ions and chemicals.

Substrate specificity and kinetic parameters of endo-xylanase





K_m = 14,56 mg/ml

V_{max} = 0,86 µmol/phút

Hoạt tính riêng đạt 171,56 IU/mg protein

4. Lytic polysaccharide mono-oxygenase (LPMO)

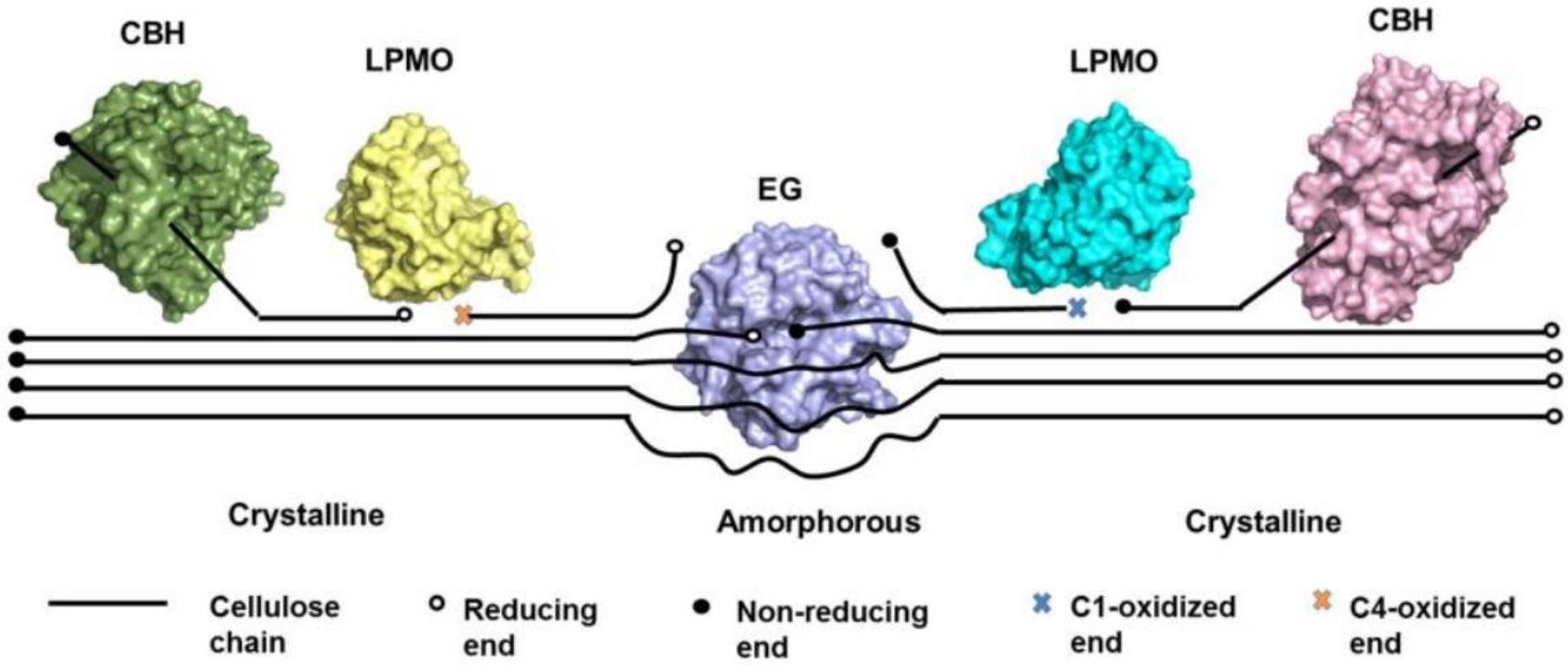


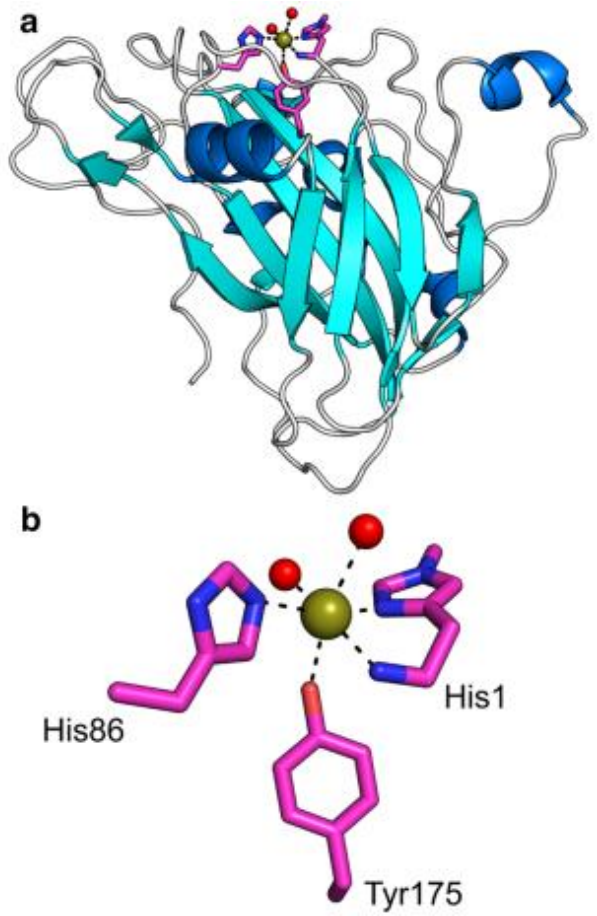
Figure 7. Schematic illustration of the cooperation between LPMOs and known cellulases (EG: endoglucanase and CBH: cellobiohydrolase) in cellulose degradation.

Bing Liu. Dissecting function and catalytic mechanism of fungal lytic polysaccharide monooxygenases. Doctoral thesis, Swedish University of Agricultural Sciences Uppsala 2019.

Identification and Characteristics of LPMOs

- **Discovery:** First reported in bacteria (2005) and fungi (2007).
- **Classification:** Belong to *Auxiliary Activity (AA)* enzyme families.
- **Families Identified: AA9, AA10**, AA11, AA13, AA14, AA15, AA16.
- **Catalytic Activity:** Oxidizes C1, C4, and C6 positions of glycosidic linkages in polysaccharide chains.
- **Substrates:** Chitin, cellulose, and starch are primary substrates.
- **Active Site Motif:** H1-Hx-F/Y; located on the enzyme surface.
- **Annotation:** Not detected by SBS (BGI) in humus samples; identified using profile HMM (69 putative LPMOs, 31 complete).

3D structure of a typical LPMO and its active site



*Ta*LPMO9A of the fungus*Thermoascus aurantiacus*
(Eijsink *et al. Biotechnol Biofuels (2019) 12:58*
https://doi.org/10.1186/s13068-019-1392-0)

Validation Workflow of Predicted LPMO Genes

1. Conserved catalytic site identification (H1–Hx–Fz) via sequence alignment
2. Phylogenetic Analysis: Construct phylogenetic trees to assess evolutionary relationships with known LPMO families.
3. Functional domain annotation: InterPro, Pfam
4. 3D structure prediction: AlphaFold2
5. LPMO activity validation: expression & assay

Amino acid sequence alignment of 31 putative LPMOs

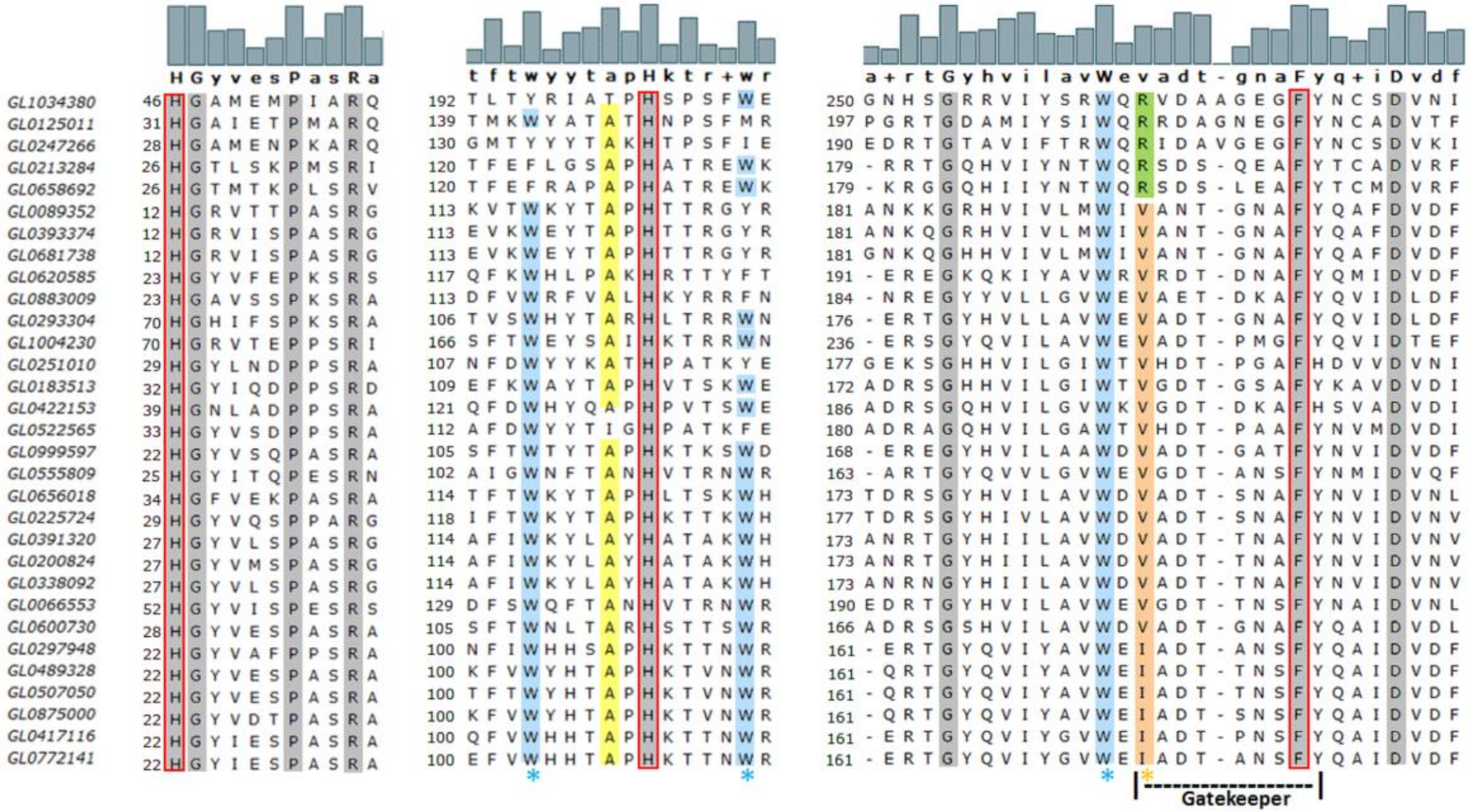


Figure 2 **Amino acid sequence alignment of regions containing three conserved amino acids in the H1-Hx-Fy motif of 31 putative LPMOs.** Gray highlighting positions indicate conserved amino acid

Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

Phylogenetic tree of 31 putative LPMOs (Cat. Domain)

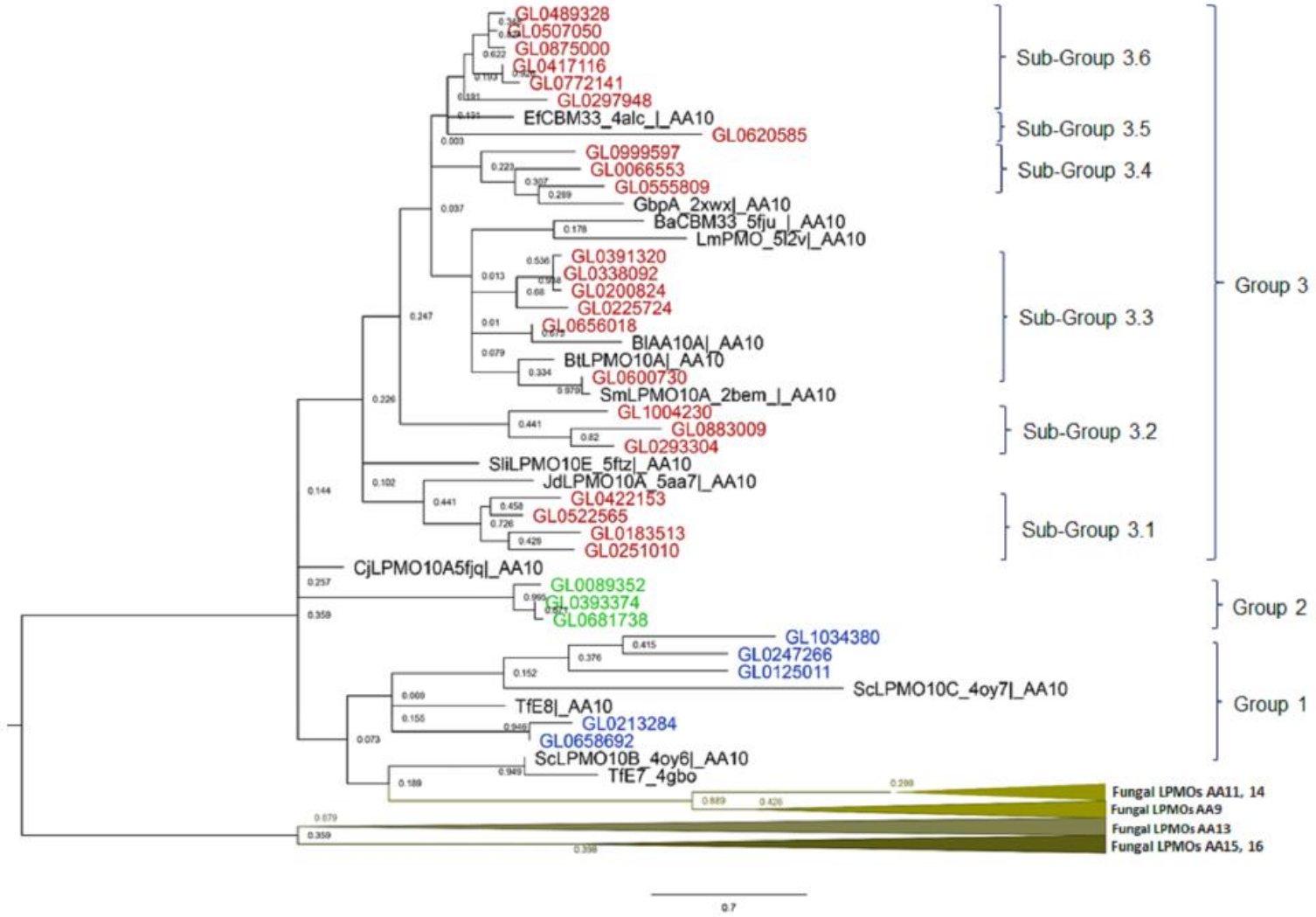
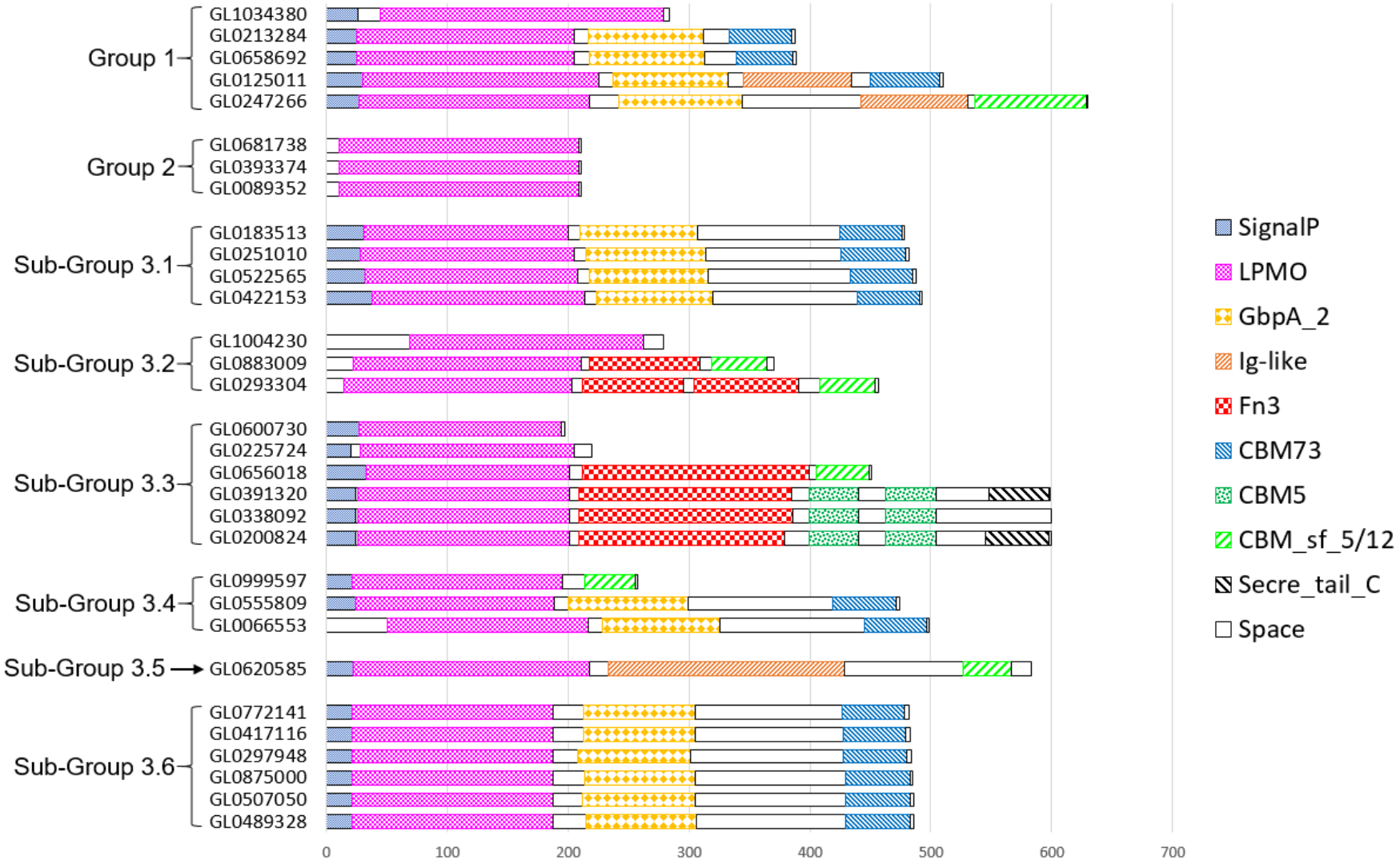


Figure 3 **Phylogenetic tree of catalytic domains of 31 putative LPMOs, 14 bacterial LPMOs, and 25 fungal LPMOs by neighborhood-joining method MEGA-X.** The bootstrap value was set to 1,000. Blue:

Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

Predicted functional domain structures of 31 putative LPMOs by InterProScan



Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

Structural characteristics of 31 putative LPMO

PeerJ

Table 1 Structure characteristics of 31 putative LPMOs.

Groups	Proteins	Length of full/mature proteins (aa)	Positions of H1, Hx, Fy in catalytic domain	Number of cysteines in full-length protein/catalytic domain
Group 1	GL0247266	631/604	H_1-H_{112}-F_{185}	4/4
	GL1034380	284/240	H_1-H_{156}-F_{227}	4/3
	GL0125011	510/481	H_1-H_{118}-F_{189}	6/4
	GL0658692	389/365	H_1-H_{104}-F_{174}	8/6
	GL0213284	388/364	H_1-H_{104}-F_{174}	8/6
Group 2	GL0393374	211/201	H_1-H_{111}-F_{191}	1/1
	GL0681738	211/201	H_1-H_{111}-F_{191}	1/1
	GL0089352	211/201	H_1-H_{111}-F_{191}	1/1
Sub-Group 3.1	GL0183513	478/448	H_1-H_{87}-F_{162}	6/4
	GL0251010	483/456	H_1-H_{88}-F_{170}	6/4
	GL0522565	488/457	H_1-H_{126}-F_{169}	6/4
	GL0422153	493/456	H_1-H_{92}-F_{169}	6/4
Sub-Group 3.2	GL1004230	279/211	H_1-H_{106}-F_{187}	4/4
	GL0883009	370/349	H_1-H_{99}-F_{182}	2/1
	GL0293304	457/443	H_1-H_{100}-F_{181}	1/0
Sub-Group 3.3	GL0656018	451/419	H_1-H_{90}-F_{161}	2/2
	GL0225724	220/193	H_1-H_{99}-F_{170}	0
	GL0391320	599/574	H_1-H_{97}-F_{168}	8/0
	GL0200824	600/575	H_1-H_{97}-F_{168}	7/0
	GL0338092	600/575	H_1-H_{97}-F_{168}	8/0
	GL0600730	197/171	H_1-H_{87}-F_{160}	4/4
Sub-Group 3.4	GL0066553	499/449	H_1-H_{87}-F_{160}	6/4
	GL0555809	474/451	H_1-H_{87}-F_{159}	6/4
	GL0999597	258/238	H_1-H_{93}-F_{167}	4/4
Sub-Group 3.5	GL0620585	583/562	H_1-H_{104}-F_{189}	4/2
Sub-Group 3.6	GL0297948	484/464	H_1-H_{88}-F_{160}	6/4
	GL0875000	485/465	H_1-H_{88}-F_{160}	6/4
	GL0489328	486/466	H_1-H_{88}-F_{160}	6/4
	GL0507050	486/466	H_1-H_{88}-F_{160}	6/4
	GL0417116	483/463	H_1-H_{88}-F_{160}	6/4
	GL0772141	482/462	H_1-H_{88}-F_{160}	6/4

Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

3D Structural comparison of active sites and catalytic domains in eight representative LPMOs

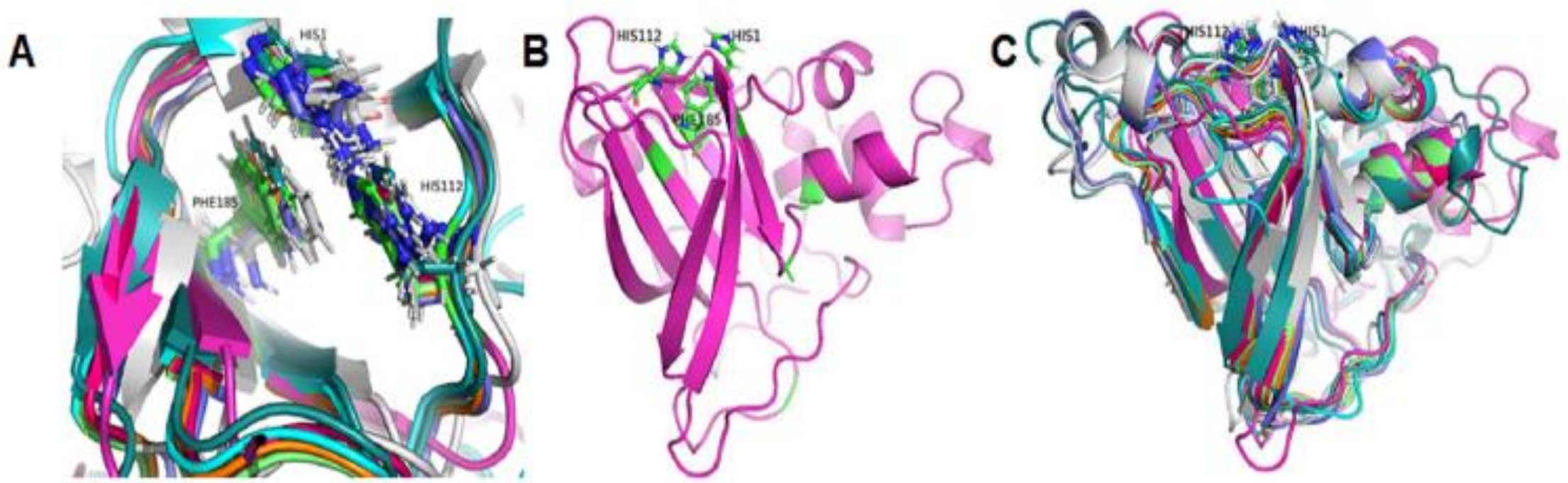


Figure 5 **Three-dimensional structure prediction of catalytic domains and spatial arrangement of conserved histidine brace and phenylalanine in active site of putative LPMOs by Alphafold2.** (A) Spatial

Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

Expression and purification of proteins GL0247266 và GL0183513

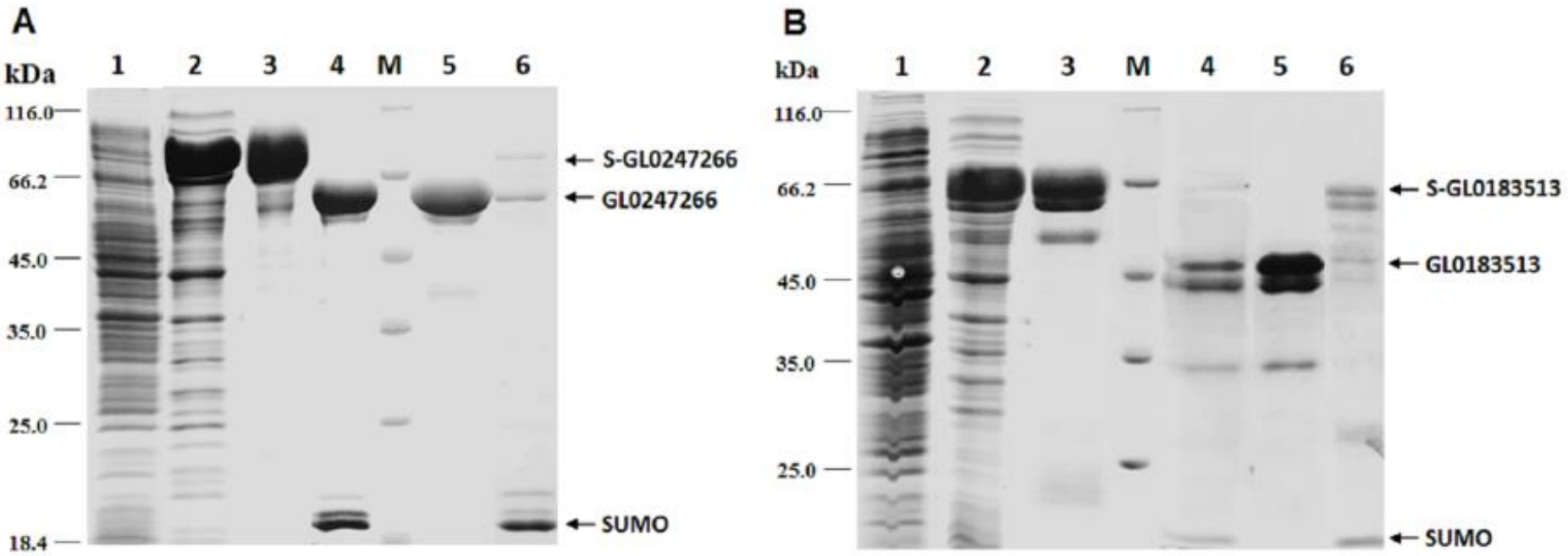


Figure 6 **Results of protein expression in *E. coli* by vector pET-SUMO3 and protein purification using His-tag affinity chromatography.** 1: Non-induced total protein fraction; 2: Induced total soluble protein fraction by IPTG; 3: Fused SUMO-GL0247266 (A) and SUMO-GL0183513 (B); 4: Cleavage products of SUMO-GL0247266 (A) and SUMO-GL0183513 (B) treated with SUMO protease; 5: Protein GL0247266 eluted at 90 mM imidazole (A) and GL0183513 eluted at 100 mM imidazole (B). 6: SUMO fraction eluted at 500 mM imidazole; M: Protein molecular mass standards (Fermentas).

Full-size DOI: 10.7717/peerj.17553/fig-6

Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

Assessment of LPMO activity of GL0183513 and GL0247266 on chitin

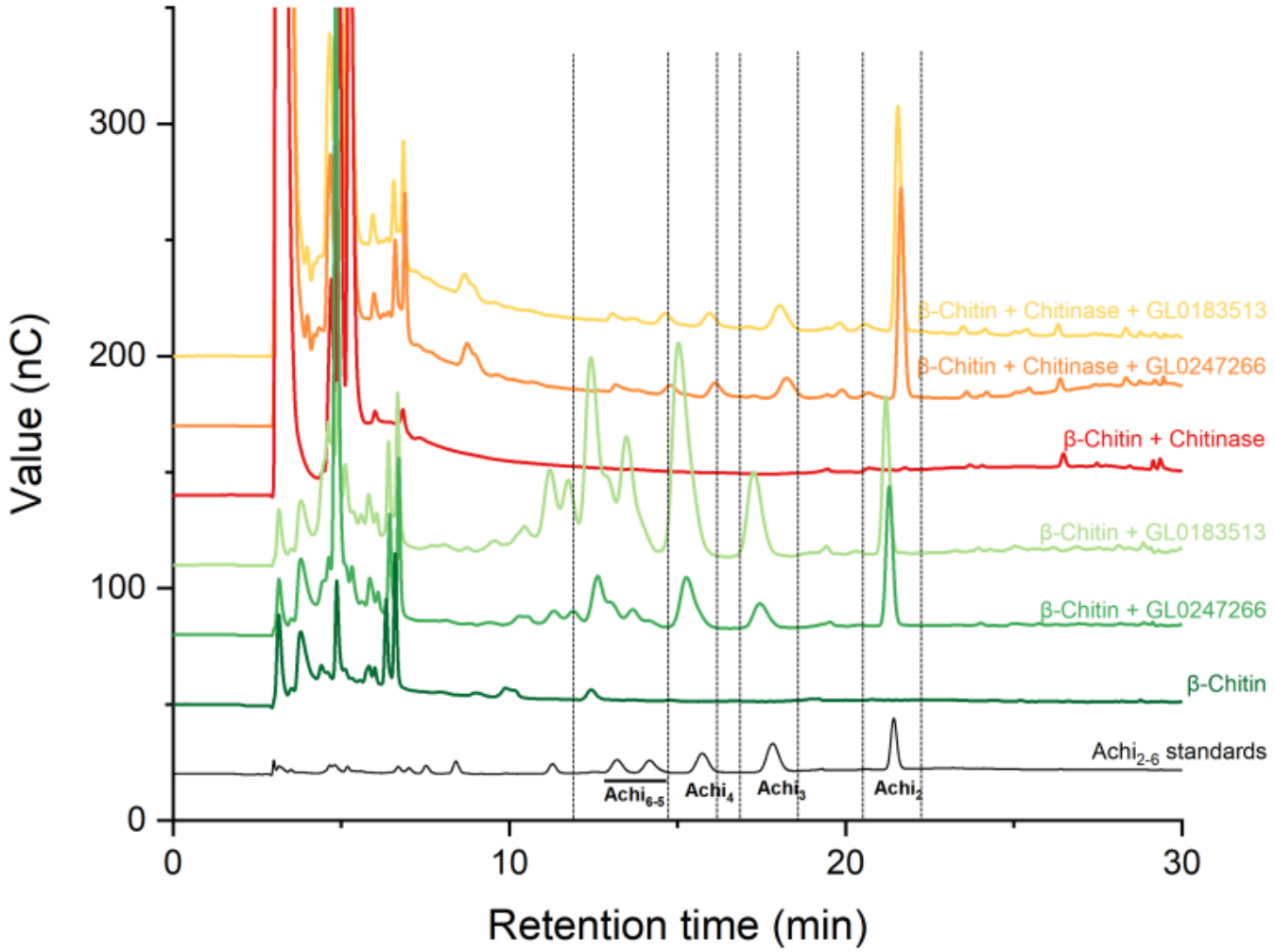


Figure 7 **HPAEC-PAD chromatograms for LPMO activity assay with *β*-chitin of GL0183513 or GL0247266 alone (green) and in combination with chitinase (red).** Key products derived from LPMO

Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

3D Structural prediction of GL0183513 and GL0247266 proteins using AlphaFold2

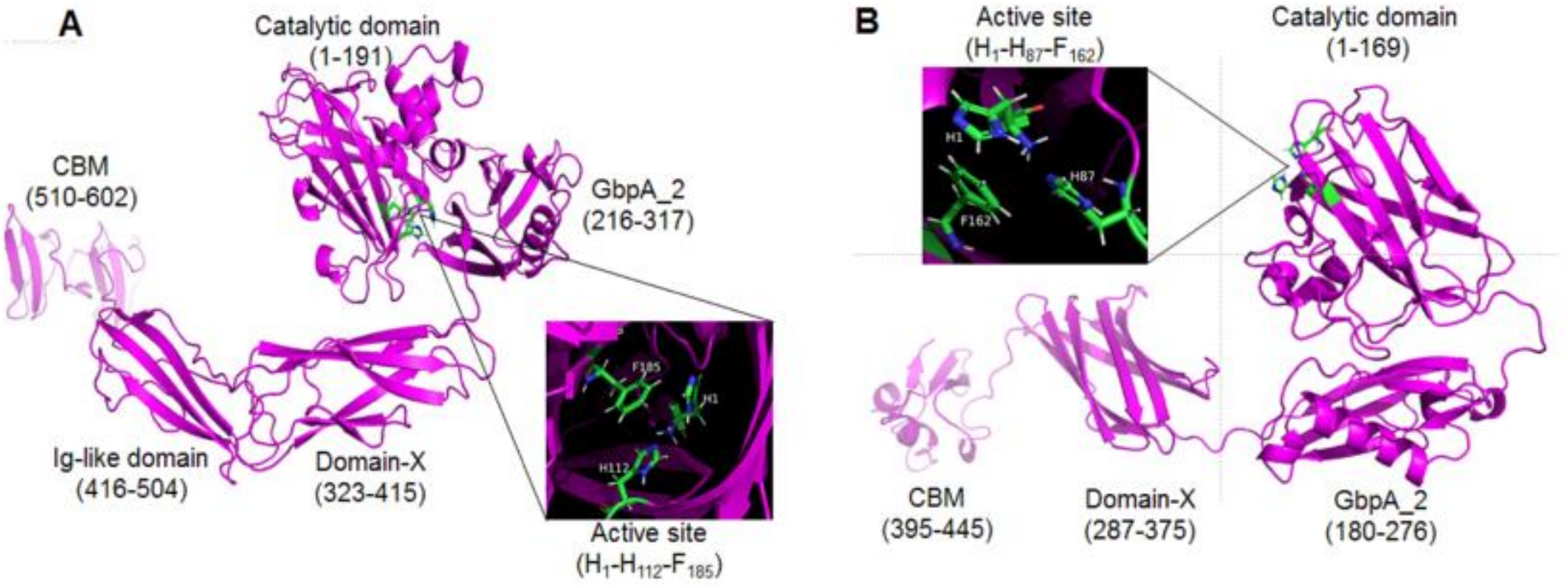
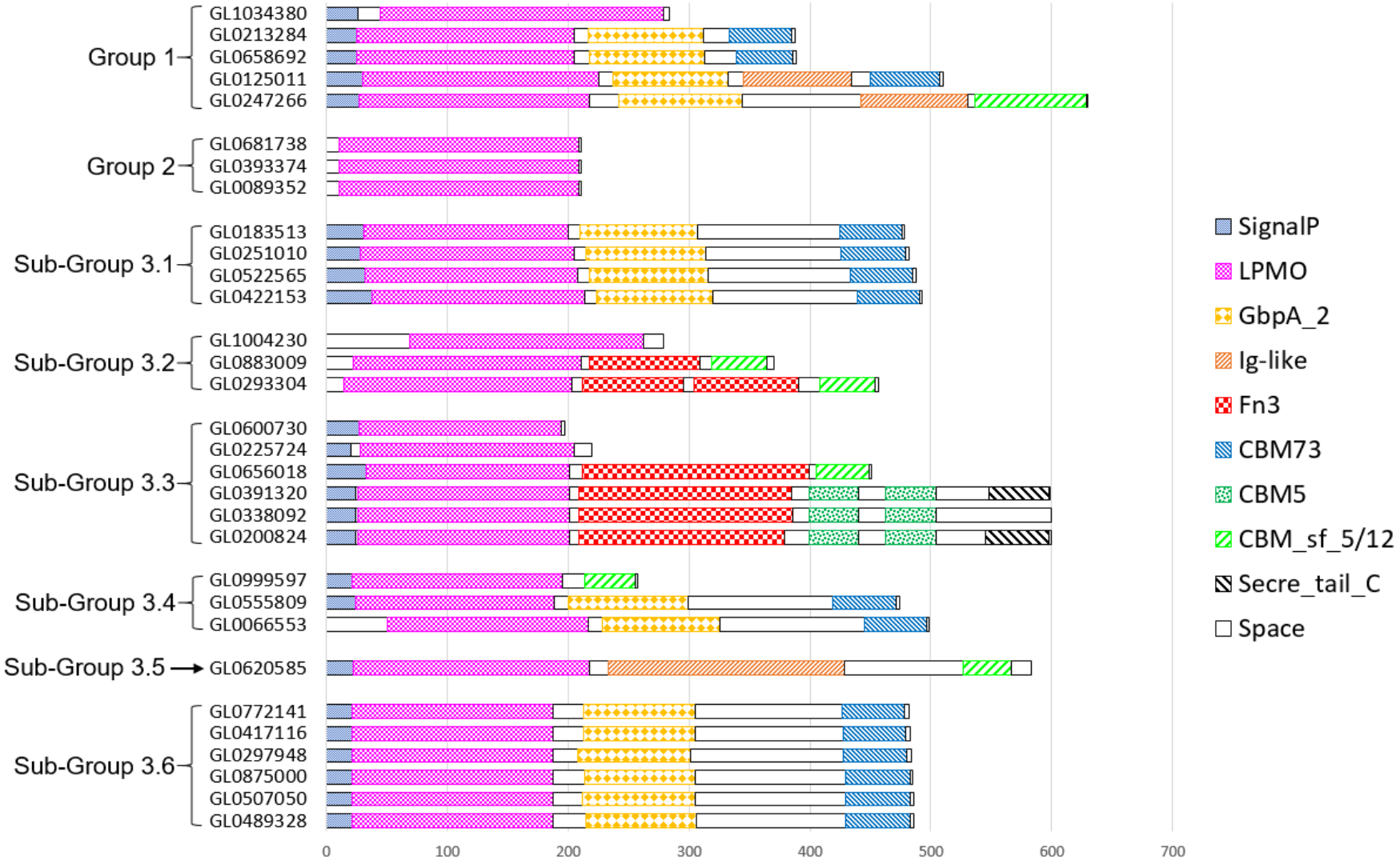


Figure 8 Three-dimensional structures of mature GL0247266 (A) and GL0183513 (B) predicted using Alphafold2. Functional domain locations are marked. The structure of the active site is zoomed in in the boxes.

Full-size DOI: 10.7717/peerj.17553/fig-8

Predicted functional domain structures of 31 putative LPMOs by InterProScan



3D Structural Comparison of Domain-X from 13 LPMOs and the GbpA_3 Domain of *Vibrio cholerae* GbpA

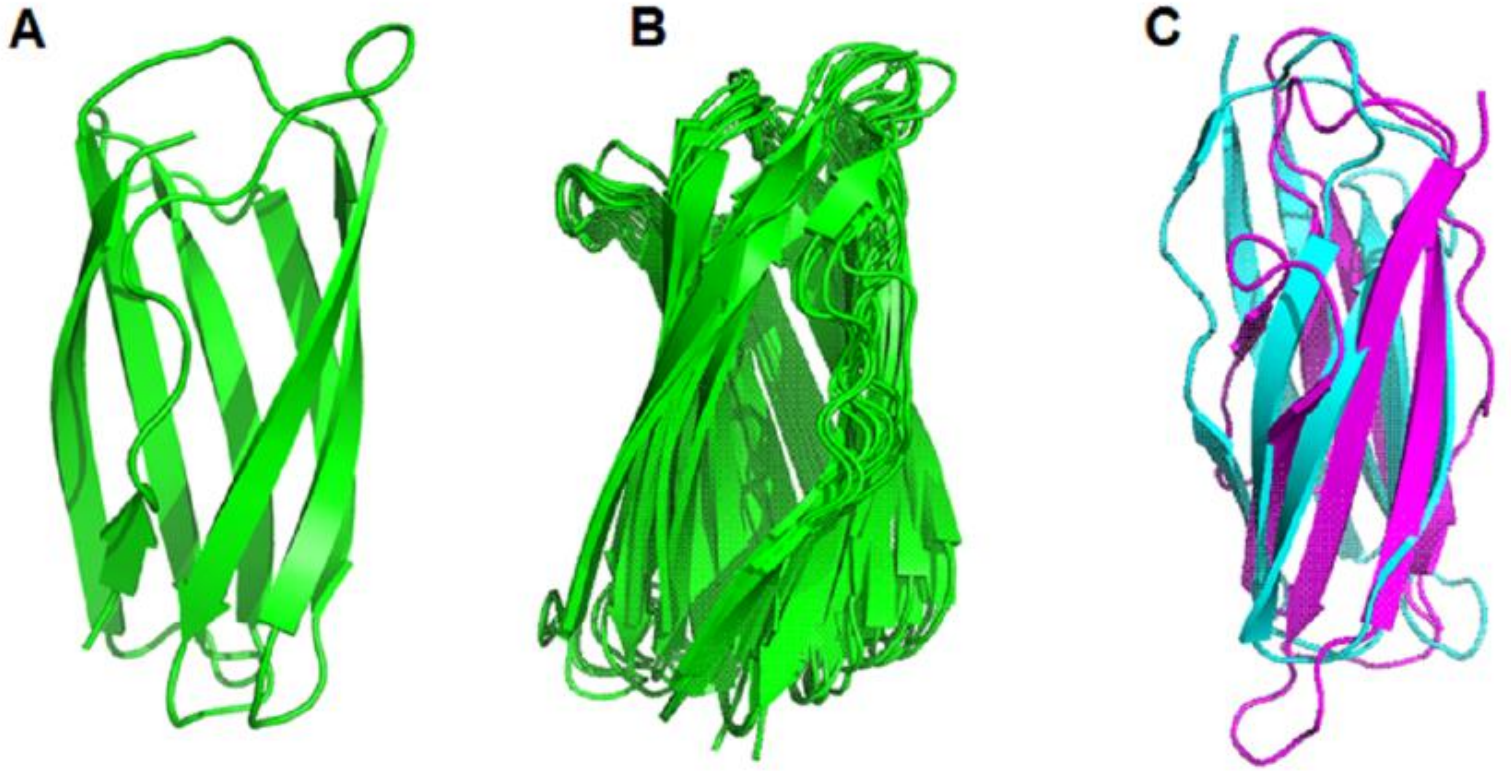
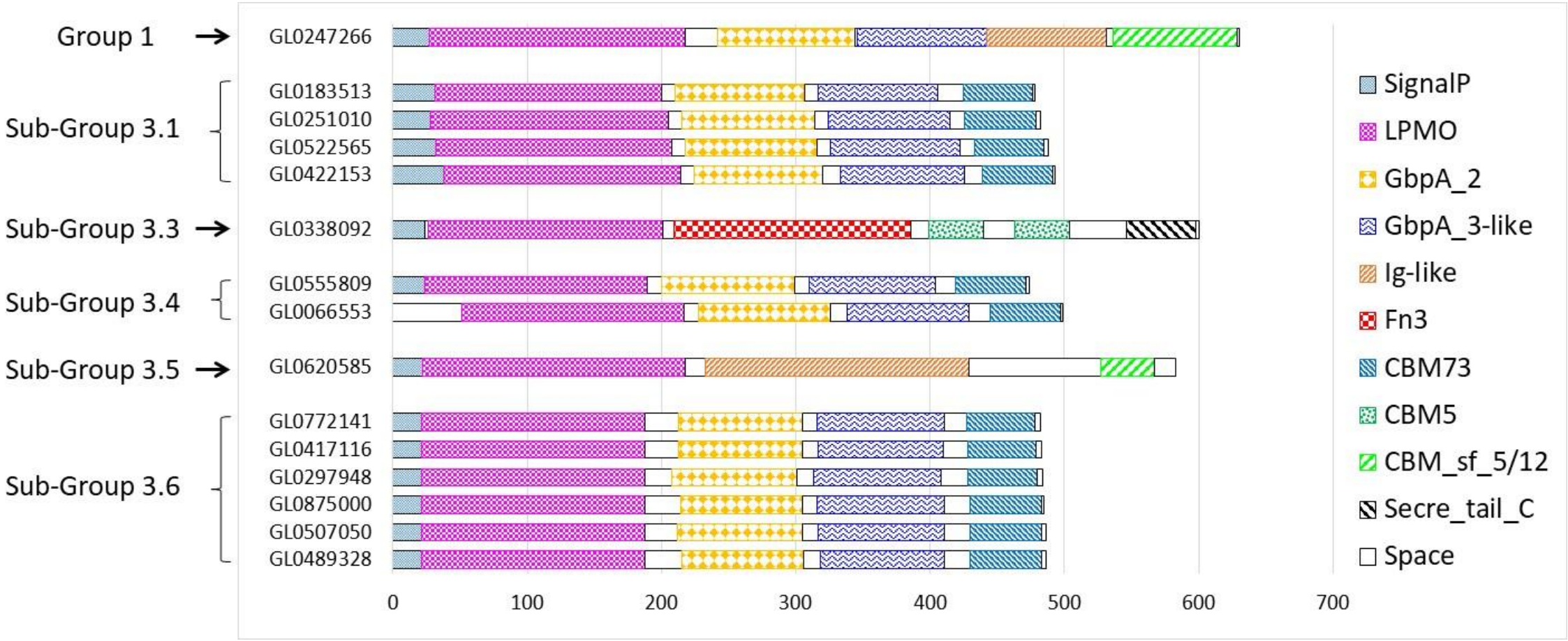


Figure 9 **Comparison of the spatial structure of the Domain-X of 13 putative LPMOs with the GbpA_3 domain of *Vibrio cholera* GbpA.** (A) Three-dimensional structure of Domain-X of protein GL0251010. (B) Spatial structure comparison of Domain-X of 13 proteins GL0251010, GL0247266, GL0422153, GL0183513, GL0522565, GL0555809, GL0066553, GL0297948, GL077214, GL0417116, GL0875000, GL0489328, and GL0507050. (C) Spatial structure comparison of GL0183513 Domain-X (magenta) with the GbpA_3 domain of *V. cholerae* GbpA (Blue).

Full-size DOI: 10.7717/peerj.17553/fig-9

Predicted functional domain structures of 31 putative LPMOs by Alphafold2



Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

5. Discussion

- **Comparative analysis of tools for mining metagenomic DNA data: SBS vs HMM**
- Possible functional roles of GbpA-like LPMOs in fungi-bacteria relationship

4. Discussion

- Comparative analysis of tools for mining metagenomic DNA data: SBS vs. HMM
- **Possible functional roles of GbpA-like LPMOs in fungi-bacteria relationship**

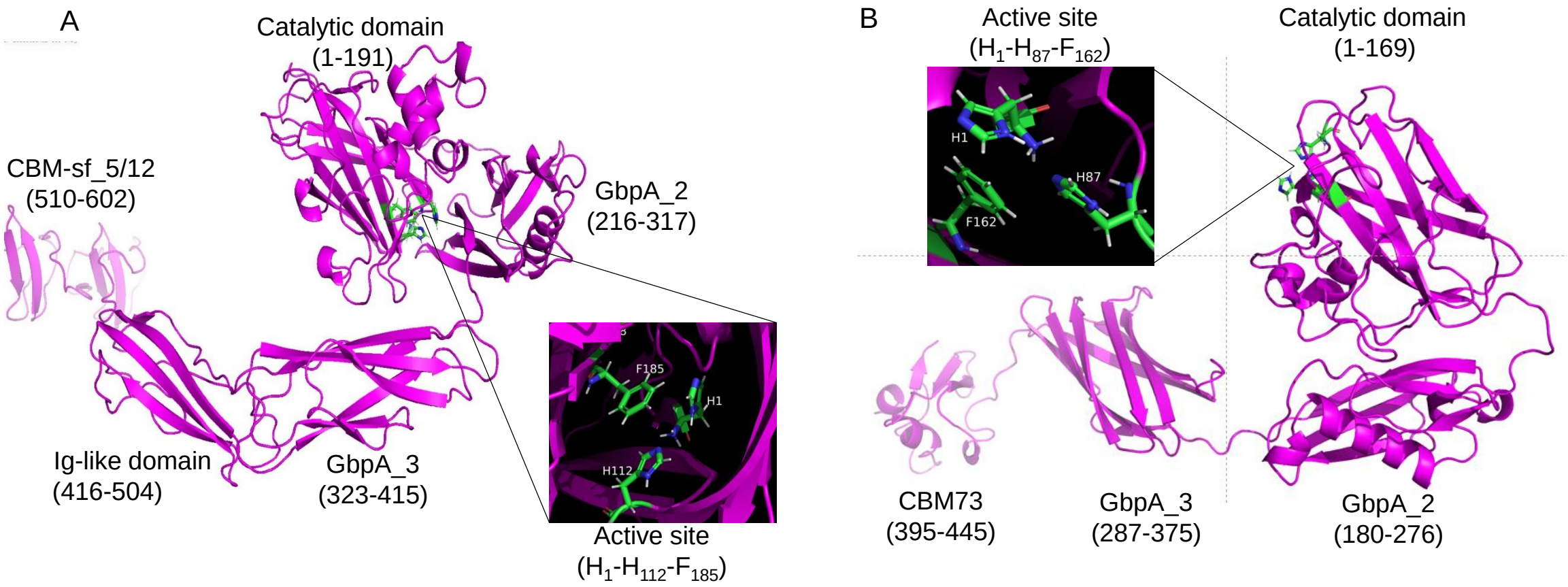
Bacteria–Fungi Interactions on Plant-Derived Substrates

Co-occurrence of bacteria and fungi can trigger symbiotic interactions:

- ❖Cooperation
- ❖Mutualism
- ❖Nutrient competition
- ❖Antagonism

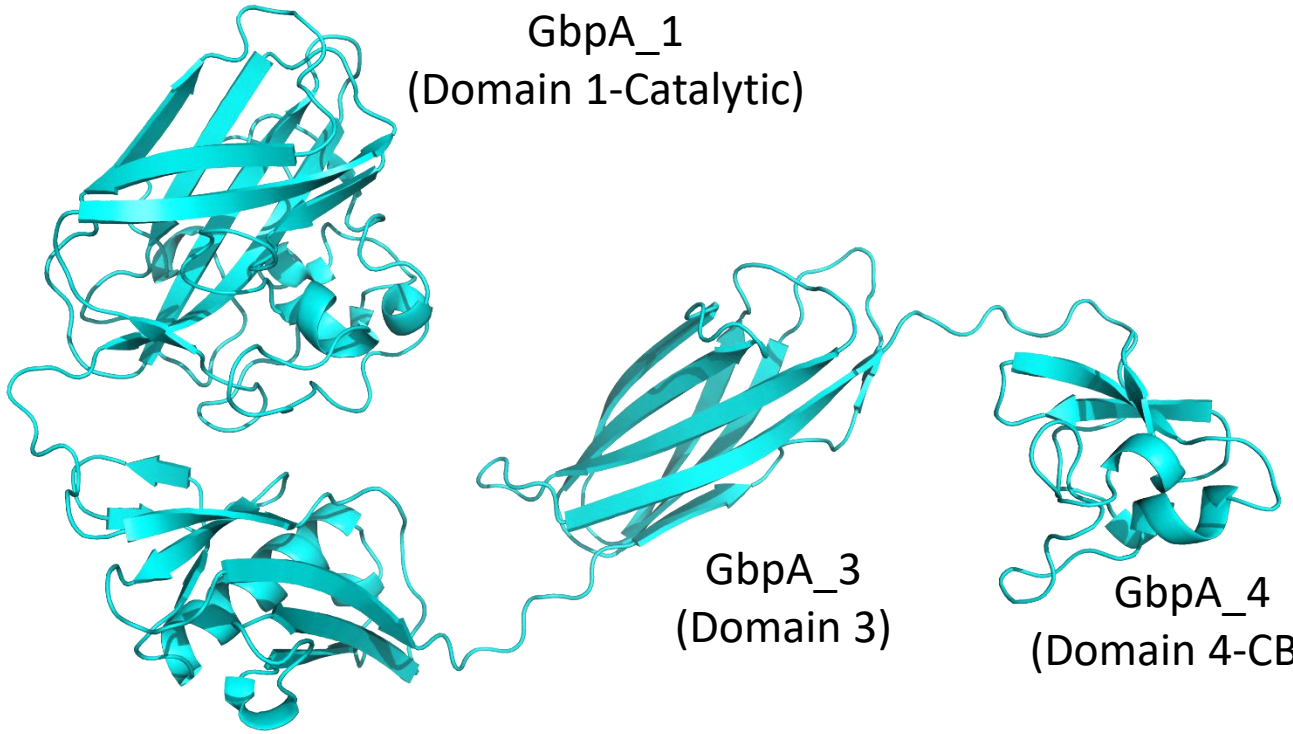
Three-dimensional structure of mature GL0247266 (A) and GL0183513 (B) predicted by the Alphafold2

(Functional domains positions are noted. Structure of active sites are zoomed in separated boxes)

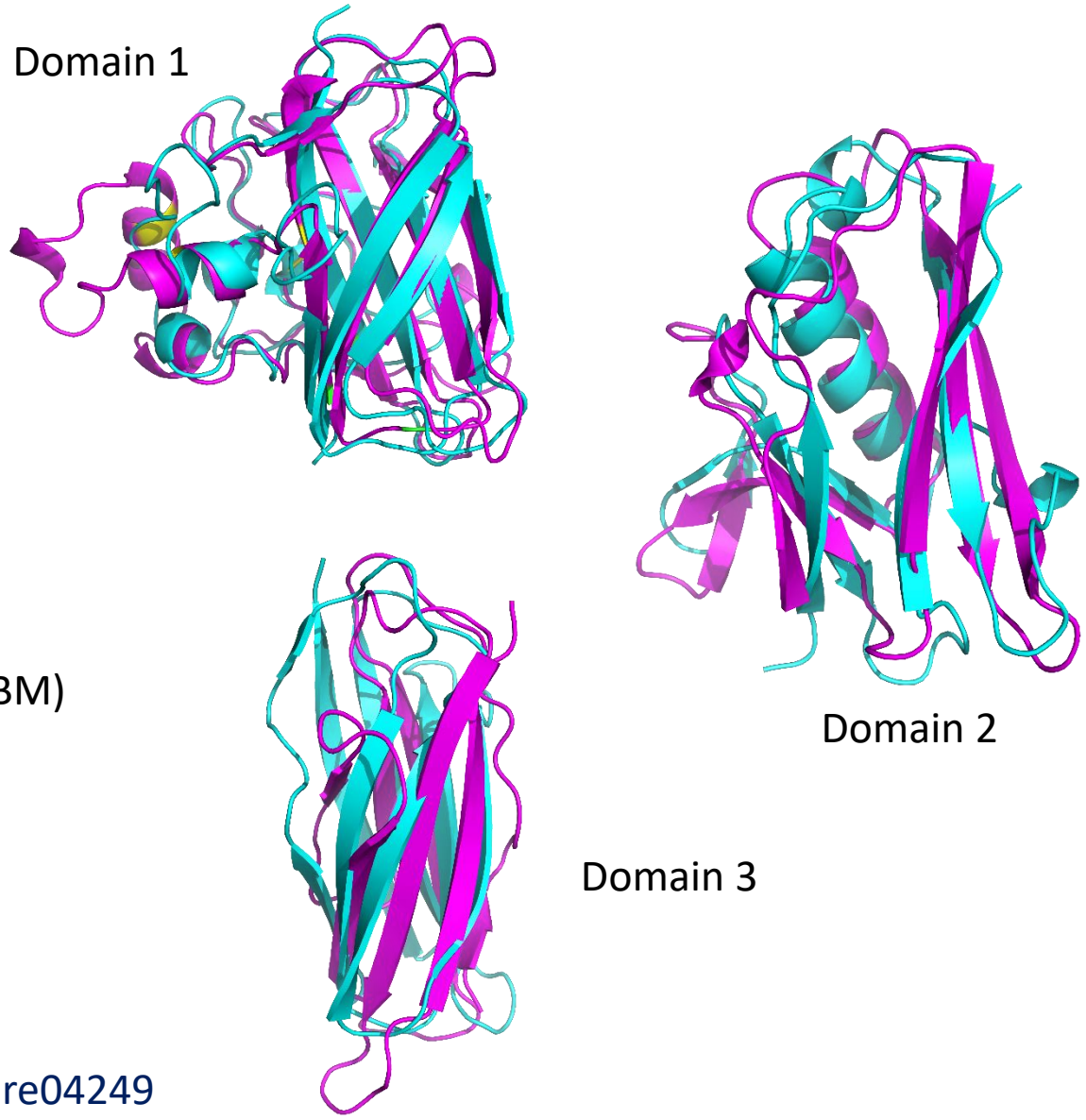


Ig-like domains are involved in a variety of functions, including cell-cell recognition, cell-surface receptors, muscle structure and the immune system [PUBMED:10698639].

GL0183513 vs GbpA of *Vibrio cholera*
(Magenta is GL0183513)

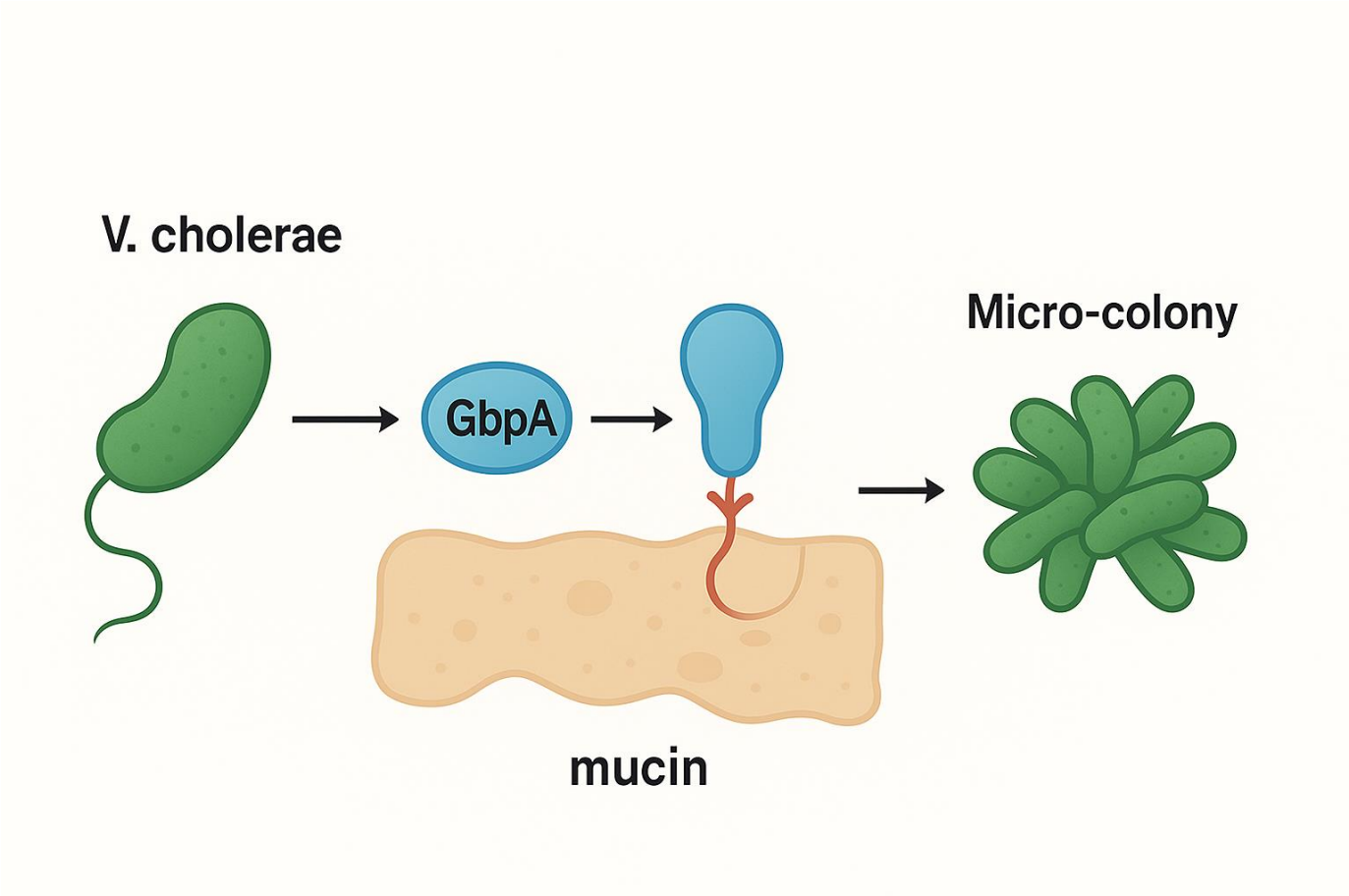


GbpA of *V. cholera*



Kirn, Jude & Taylor, 2005. *Nature* **438**:863–866. DOI 10.1038/nature04249

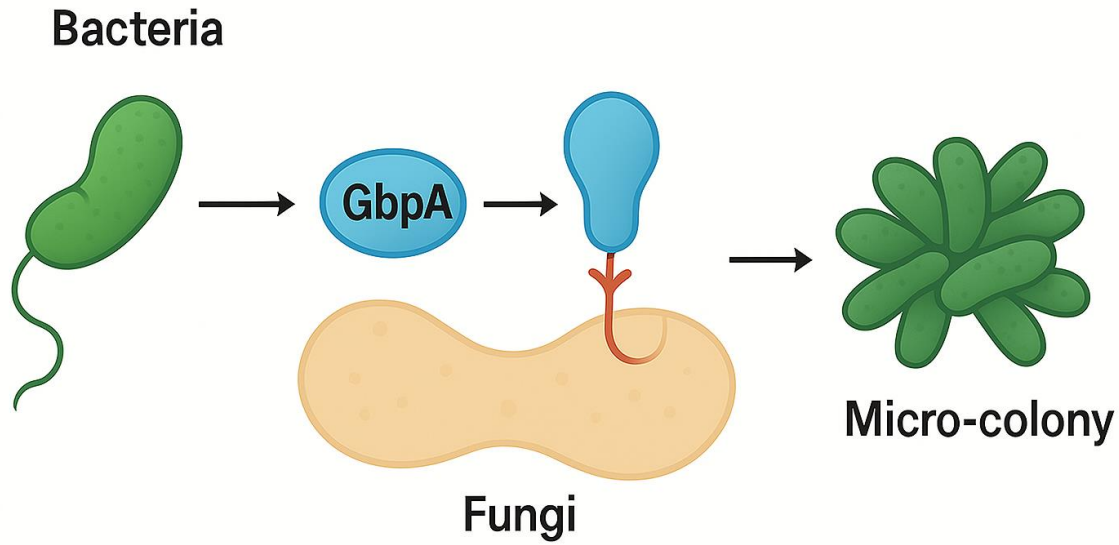
Mechanism of GbpA-mediated colonization in *Vibrio cholerae*



1. Bacteria *V. cholerae* secreted GbpA.
2. GbpA attaches domain 1 and 4 to mucin (intestinal surface), marks mucosal surface for colonization.
3. Domains 2–3 bind back to bacteria surface → Facilitate bacterial adhesion and micro-colony formation.

Kirn, Jude & Taylor, 2005. *Nature* **438**:863–866. DOI 10.1038/nature04249

Hypothesis of mechanism of GbpA-like - mediated bacterial colonization on *fungi* surface



1. Bacteria secreted GbpA.
2. GbpA attaches domain 1 and 4 to fungi surface, marks fungal surface for colonization.
3. Domains 2–3 bind back to bacteria surface → Facilitate bacterial adhesion to fungal surface and micro-colony formation.

Truong et al. (2024), PeerJ, DOI 10.7717/peerj.17553

Acknowledgments

- MOST of Vietnam and BMBF of Germany for research funds:
  - Prof. Jurgen Pleiss (Stuttgart Univ.) and Prof. Wolfgang Streit (Hamburg Univ.) Germany. (2018-2021).
  - Prof. Nico M. van Straalen and Dr. Dick Roelofs, Vrije Universiteit Amsterdam, the Netherlands. (2014-2017).
  - Dr. Keitarou Kimura, National Food Research Institute, Tsukuba, Japan (2012-2014).
- VAST project (2011-2012).
- Institute of Biotechnology, VAST.
- Researchers of GEL, IBT.
- Dr. Vu Van Van, Hi-Tech Institute, NTT University, HCM City.
- Dr. Nguyen Hong Thanh, Vimec Hi-tech Center, Vinmec Healthcare system.

Thanks for your attention!