Universiteit
Leiden
The Netherlands

# Microbiome-mediated colonization resistance: defense against enteropathogens and multi-drug resistant organisms

Ducarmon, Q.R.

**Citation**
Ducarmon, Q. R. (2022, March 23). *Microbiome-mediated colonization resistance: defense against enteropathogens and multi-drug resistant organisms*. Retrieved from https://hdl.handle.net/1887/3280022

| | |
|---|---|
| Version: | Publisher's Version |
| License: | Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden |
| Downloaded from: | https://hdl.handle.net/1887/3280022 |

**Note:** To cite this publication please use the final published version (if applicable).

# Chapter 5

## Development of a novel computational tool for profiling of carbohydrate-active enzymes (CAZymes) in the human gut and its application in colorectal cancer cohorts

Unpublished data, manuscript in preparation

**Q. R. Ducarmon[a,b,c,*], N. Karcher[a,*], H.L.P. Tytgat[d], C. Schudoma[a], G. Zeller[a]**

[a]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.
[b]Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands.
[c]Center for Microbiome Analyses and Therapeutics, Leiden University Medical Center, Leiden, The Netherlands.
[d]Nestlé Institute of Health Sciences, Nestlé Research, Société des Produits Nestlé S.A., Lausanne, Switzerland
*these authors contributed equally

# Abstract

Carbohydrate-active enzymes (CAZymes) are essential for the synthesis and breakdown of (complex) glycans and glycoconjugates. They are present in all living species, but are especially diverse in bacteria. In the gut microbiome, CAZymes are crucial for metabolizing complex carbohydrates of dietary and host origin, such as fiber and mucins, respectively. Currently, dbCAN2 is the most widely used computational tool for annotation of CAZymes in genomic data, but it cannot be directly applied to metagenomic data. dbCAN2 can identify protein sequences that are similar to those present in the CAZy database (the most comprehensive data and knowledge base on CAZymes) using Hidden Markov models (HMMs) specifically built for each CAZyme (sub-)family. However, detection accuracy (E-values cutoffs) has not been calibrated for these HMMs. A second challenge for wide application of this tool for metagenome analysis is the lack of systematic substrate annotations for CAZyme families, which are currently primarily grouped based on amino acid sequence similarity. A hierarchical annotation of CAZyme substrates would however be needed for functional interpretation of CAZyme profiles. To close this gap, the main aim of this study was to build the first tool for computing CAZyme profiles from shotgun metagenomic data, which can be interpreted in terms of substrate specificities. This entailed optimization of HMM E-values for precise detection of CAZymes and construction of a novel hierarchical substrate scheme to facilitate functional interpretation. Application of this tool using data from eight different colorectal cancer (CRC) cohorts revealed that CRC metagenomes were enriched in microbial CAZymes involved in glycosaminoglycan metabolism (p-value 7.44e^-04) and in peptidoglycan metabolism (3.52e^-02), and depleted in CAZymes involved in dietary fiber metabolism (p-value 3.68e^-04) as compared to control metagenomes, suggesting that known dietary risk factors, such as increased meat consumption/decreased fiber consumption in CRC, are reflected in the gut microbial CAZy repertoire.

# Introduction

Carbohydrate-active enzymes (CAZymes) are a diverse group of enzymes which can build up and break down glycans and glycoconjugates, consisting of e.g. glycosyltransferases and glycosyl hydrolases. They are present in all domains of life, but an exceptionally diverse set is encoded in microbial genomes, thereby contributing to the extraordinary metabolic versatility of microorganisms[1]. Microbes present in the human gut possess a variety of CAZymes which, amongst others, aid in metabolizing (complex) carbohydrates consumed from the diet, one of the key functions of the gut microbiome[2]. As the human genome encodes for only 17 enzymes with limited capacity (involved in breakdown of sucrose, lactose and starch) for carbohydrate degradation[1], microbial CAZymes are imperative for metabolizing diet-derived complex carbohydrates.

Fermentation of these complex carbohydrates leads to production of metabolites including short-chain fatty acids (SCFAs), of which butyrate, propionate and acetate form 90-95% of the total SCFA pool in the human gut[3]. The contributions of SCFAs to maintaining host health are difficult to overstate, but amongst the most important functions are butyrate being the main energy source for enterocytes[4], regulating gut barrier integrity[5] and being ligands for a variety of receptors present on enteroendocrine and immune cells[5].



Apart from metabolizing complex carbohydrates that are stemming from the diet, CAZymes can also metabolize host glycans such as the mucus layer, with negative consequences in certain conditions. For example, upon exposing fiber-deprived mice to the pathogen *Citrobacter rodentium*, mice suffered from a far higher mortality than mice who were provided a normal diet. This was likely due to the microbiome starting to use host-derived glycans (mucus layer) as the main energy source due to the lack of fiber. This led to mucus layer erosion and thereby to closer proximity of *C. rodentium* to the gut epithelium in these fiber-deprived mice, which likely facilitated infection[6].

In the CAZy database, CAZyme families have been constructed and they are defined based on amino acid sequence similarity. For each family, at least one founding member is required to have been biochemically characterized[7]. These CAZyme families are rigorously curated and extensively evaluated before a new CAZyme family is accepted into the database[7]. This is one of the reasons why the CAZy database is regarded as the gold standard in the field. One drawback of the current classification scheme, based on amino acid sequence similarity, is that glycan substrate annotation is in many cases problematic. A CAZyme family can contain enzymes with similar amino acid sequences, while they can metabolize a variety of substrates[8]. This drawback can, however, not only be ascribed to the current classification scheme. For multiple CAZyme families

no substrate is known and to this day it remains unknown what determines substrate specificity. Nevertheless, the lack of a higher, systematic substrate system for CAZyme families for which substrates are known currently limits functional interpretability with regard to high-level substrate usage.

A main goal of researchers interested in CAZymes is to discover which CAZymes are encoded in their (meta)genomic data and whether CAZyme abundances differ meaningfully between groups of samples. One of the most popular tools to annotate CAZymes in genomic data is the dbCAN2 tool, which uses CAZyme (sub)family-specific Hidden Markov models (HMMs) to annotate gene sequences[9]. While this tool has proven valuable and remains widely used, optimization of E-value cutoffs (which aims to minimize conflicts between functional homology and sequence similarity) was performed on only six genomes[9]. This leaves considerable room for optimizing the detection accuracy of CAZyme annotation. In addition, dbCAN2 is an open-reading frame (ORF) annotation framework and not a profiler, which prevents estimating taxonomic or functional abundance profiles from a shotgun metagenome. From the ORF prediction it is not trivial to estimate (relative) abundances of CAZymes in a given metagenome, especially given that, due to their multimodular nature, multiple CAZyme families can occur within a single ORF. Lastly, to be able to annotate ORFs with CAZymes, researchers would first have to laboriously reconstruct these (via metagenome assembly, ORF prediction, and redundancy removal) from metagenomic data, posing a substantial challenge to many (non-)bioinformaticians.

Here, we aimed to develop the first easy-to-use tool to profile CAZymes from short-read metagenomic data and to optimize E-values for more accurate detection of CAZymes using HMMs. In addition, we suggest a novel substrate annotation scheme allowing for improved interpretation of the resulting CAZyme profiles in terms of broader substrate groups these act on. Lastly, we applied our tool on eight different colorectal cancer (CRC) cohorts to uncover novel associations between specific CAZymes, substrate metabolism and CRC.

## Materials and methods

**Generation of CAZyme (sub)family module sequence set**

In order to obtain family-wise CAZy modules, we first downloaded all CAZy sequences (http://bcb.unl.edu/dbCAN2/download/Databases/V9/ dbCAN HMMdb release 9.0 and CAZyDB released on 07/30/2020). We then identified CAZy modules on these sequences using the dbCAN HMMs (from the same dbCAN2 release) using an E-value threshold of 1e-15 and a coverage threshold of 0.35 (default cut-offs on dbCAN server). We then

generated (sub)family-wise multiple sequence alignments of module sequences using mafft for the 29 largest families (containing most sequences) and mafft-linsi for the remaining sequences on representative sequences obtained from mmseqs2 clustered at 99% (clustered using mmseqs2 (arguments easy-cluster, –min-seq-id 0.99, --cov-mode 0, -c 0.5)[10, 11]. The clustering was done in order to avoid bias in building the HMMs due to overrepresentation of very similar sequences. Default parameters were used unless stated otherwise.

**Optimization of HMM E-value cutoffs**

To optimize HMM E-value cutoffs for each CAZy family we optimized HMM E-values using 5-fold blocked cross validation as well as an external negative sequence set: For each CAZy (sub)family and fold, we divided the module sequence set into two sets: The first set of sequences was used to build the module HMM and the second set was used as positive instances for evaluation, combined with the module sequences of all remaining CAZy (sub)families as well as non-CAZy sequences obtained from UniProt (see below) as negative instances. Division into training and test folds was done in a blocked fashion, where sequences in one block are always together in either training or test set. This was done to minimize information from the training set leaking into the test set. To define blocking groups, module sequences within each family were clustered at 60% sequence identity using mmseqs2 (arguments easy-cluster, –min-seq-id 0.60, --cov-mode 0, -c 0.5)[11]. Folds were designed in such a way that test sets never overlap with each other (Figure 1). This setup was applied to 330 CAZyme families with at least 150 sequences in the CAZy database. For the 207 families which had less than 150 sequences, but more than 30 sequences, cross validation was performed without blocking. Non-CAZy sequences were obtained from UniProt in the following manner. First, all manually curated sequences (Swiss-Prot) with an annotation score of five out of five (indicating experimental evidence at protein level) were downloaded (n=54,978 sequences, March 5th 2021). We subsequently filtered out all sequences with an annotated Enzyme Commission (EC) number present in CAZy, yielding 51,507 sequences. Since CBMs are non-catalytic (and thus have no EC number), we did not add UniProt sequences as negative instances when we optimized E-values for CBMs.

Finally, we determined family- and fold-wise optimal E-values by iterating over E-value thresholds (from 1e^0 to 1e^-200) and choosing the E-value that maximizes the F1-Score. The F1 score was calculated using the follow formula: 2*(recall*precision) / (recall + precision). Recall was calculated using the formula: TP / (TP + FN) and precision using TP / (TP + FP). A schematic overview of the workflow for building novel HMMs and optimization of their E-values can be found in Figure 1.

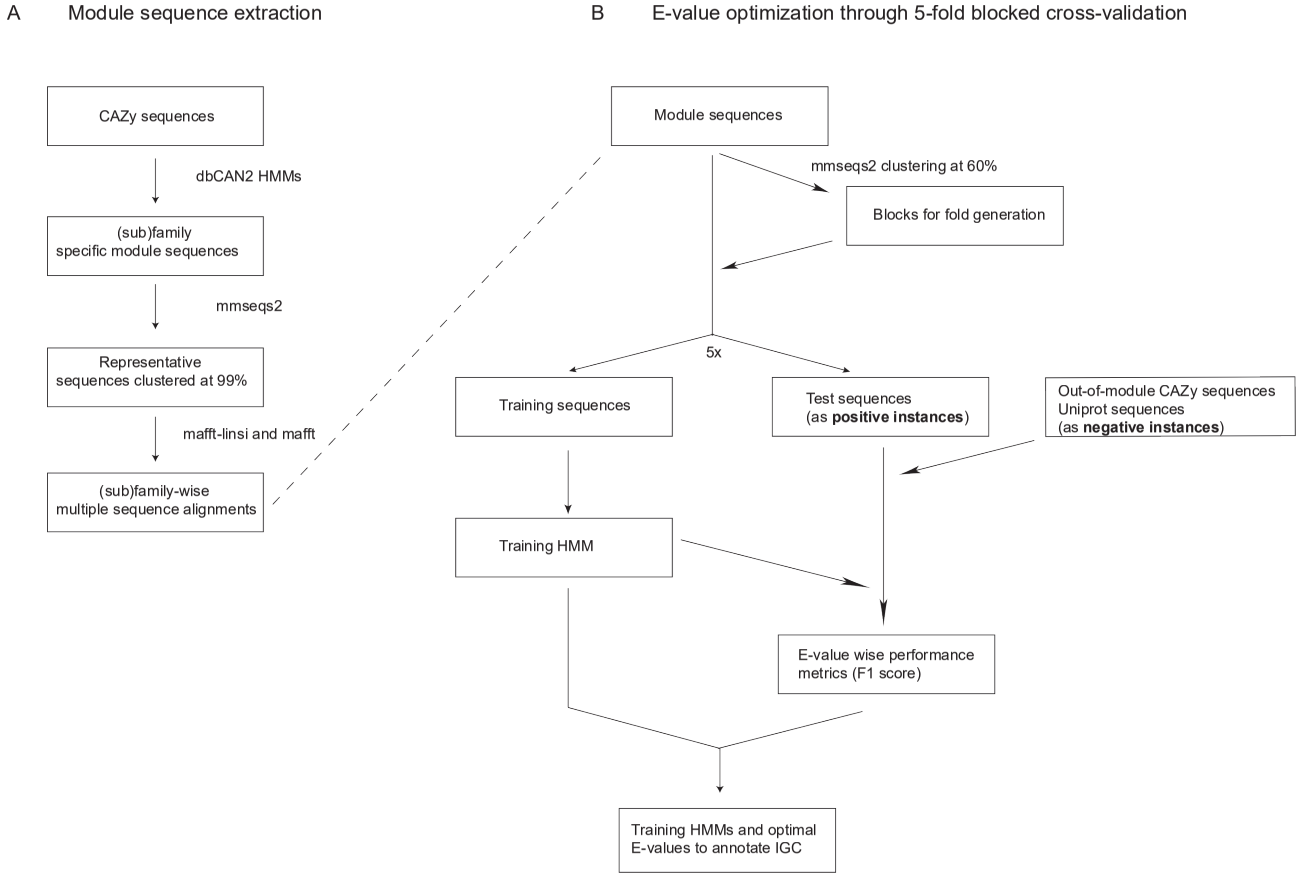


**Figure 1:** Schematic workflow of construction of the novel HMMs and optimization of their E-values. (A) Extraction of module subsequences from CAZy ORFs using the dbCAN2 HMMs. (B) E-value optimization for each of CAZy family using blocked cross-validation.

**CAZyme annotation of the Integrated Gene Catalog (IGC)**

To annotate the approximately 9.9 million genes present in the IGC[12] with CAZy module information, we ran the CV HMMs of all folds against the IGC and assigned module hits by determining the overlap of all fold-HMMs hits.

**Community profiling of datasets**

We investigated eight different colorectal cancer cohorts. Raw metagenomic sequencing data was processed using NGLess (v1.0.0) and accompanying tools[13, 14]. In short, raw sequence data was pre-processed by quality-based trimming and reads with quality value below 25 were discarded, followed by discarding reads shorter than 45 bp. Second, reads were aligned to the human genome (hg19 reference) and discarded if reads mapped with more than 90% sequence identity and an alignment length of at least 45 bp. Third, we mapped filtered reads against the Integrated Gene Catalog (IGC) using BWA-MEM and obtained BAM files were sorted using samtools sort[12, 14, 15]. Fourth, bedtools intersect (-bed -wo) was used to investigate overlap between the mapped reads and CAZy modules[16]. Lastly, we used isect_quant (https://github.com/cschu/gff_quantifier/blob/master/gffquant/isect_quant.py) to obtain CAZyme abundance profiles (which were corrected for the length of the respective CAZyme). Here, we focused on colorectal cancer (CRC) patients and controls (total n = 1225, CRC cases n = 632, controls n = 593) from eight different cohorts spanning seven countries and three continents[17-23].



**Meta-analysis of differentially abundant CAZymes in CRC**

We started our meta-analysis by investigating which CAZymes were differentially abundant between CRC and controls' metagenomes through a blocked Wilcoxon-rank sum test as available through the coin (v1.4-1) package, where study was the blocking factor[24]. Next, to investigate whether signatures were consistent across studies, univariate nonparametric Wilcoxon tests and false discovery rate (FDR) correction were performed as implemented by the SIAMCAT package (v1.10.0), on a per-study basis[25]. Generalized fold changes for both tests were calculated as implemented by the SIAMCAT package. P-values were adjusted using FDR correction (Benjamini-Hochberg method)[26] and adjusted p-values $< 0.05$ were considered significant.

**Gene set enrichment analysis (GSEA)**

Leveraging the extensive manual substrate annotations allowed us to perform GSEA to investigate whether there is differential metabolism of specific substrates between CRC patients and controls[27]. GSEA was performed using the R package fgsea(v1.16.0) and the fgseaMultilevel function with default parameters[28]. As input measure for fold change, generalized fold changes as obtained from SIAMCAT were used.

# Results

**Overview over the new CAZy profiler**

We started building our profiler by annotating the IGC with our cross-validated HMMs and optimized E-value cutoffs, obtaining CAZy module annotations and locations on each ORF. Subsequently, we mapped quality-filtered and human-filtered reads from a given metagenomic sample to the entire IGC. We then computed overlaps of the aligned reads with the annotated CAZy modules using bedtools intersect[16]. We generated CAZy profiles by averaging individual CAZy module coverages (number of reads aligning to a module normalized by its length).

**Overview of CAZy annotations on IGC**

We started our analyses by obtaining a global overview of which CAZymes were annotated using our method in the IGC, a comprehensive gene catalogue constructed from metagenomes of the human gut microbiome[12]. We detected a total of 457 unique CAZymes in the IGC, which amounted to 225,946 unique ORFs (2.29% of IGC). The largest number of (sub)families was annotated in the glycoside hydrolases (GHs) category; 240 unique (sub)families, followed by 76 glycosyltransferases (GTs), 63 carbohydrate-binding modules (CBMs) 55 polysaccharide lyases (PLs), 14 carbohydrate esterases (CEs) and 9 auxiliary activities (AA)s.



**Constructing a novel glycan substrate annotation scheme**

To construct a more informative scheme for glycan substrate annotation, we grouped carbohydrate substrates into five biological categories which we systematically annotated for each commonly encountered glycan in the human gut (Supplementary Table 1). We hereby focused on CAZy categories GH, PL, CE and CBM for two main reasons. First, there is relatively little information about the precise substrates of many GT and AA families precluding their grouping into functional categories. Second, our main interest was in digestion rather than synthesis of (complex) carbohydrates and GTs are less involved in the former process.

Based on the CAZymes that we annotated in the IGC, a list of all potential substrates was generated. Based on literature[1, 6, 29-32], these substrates were further classified into broad biological categories: reflecting their origin (e.g. bacterial, plant), function in their original biological context (e.g. storage, structural) and function at destination/general biological role (e.g. dietary fiber, glycosaminoglycan) (Supplementary Table 1). We then annotated each CAZyme detected in the IGC with substrates and matched this to the annotations per glycan (Supplementary Table 2). This way of grouping and annotating CAZymes allows for a systematic investigation of substrate metabolism potential using metagenomic data.

**Meta-analysis shows CAZymes to be differentially abundant between CRC patients and controls**

Shotgun metagenomic data from eight different CRC study populations recruited in seven different countries (Austria, China, Germany, France, Italy, Japan and the US) were re-analyzed.

The first aim of this meta-analysis was to investigate whether CRC-specific CAZyme signatures could consistently be identified across the included studies. To this end, we pooled data from all studies and performed a blocked Wilcoxon test to account for study heterogeneity (as blocking factor). As a result, we identified 203 CAZymes to be significantly differentially abundant between CRC patients and controls at an FDR-adjusted p-value $\leq 0.05$ (data not shown). Next, we aimed to identify which CAZyme signatures were consistent across the cohorts. Therefore, we explored the datasets by investigating differential abundance of CAZymes separately in each study by employing a Wilcoxon test and FDR correction. Significantly differentially abundant CAZymes were noted in all cohorts except from the US cohort, which previously showed weak associations between the microbiome and CRC, possibly related to the long-term sample storage (Figure 2)[22, 23].



Among the consistently more abundant CAZymes in controls across different cohorts were GH94, GH53, GH5_2, CE17, CBM48 and CBM2. Of these CAZymes, GH94, GH53, GH5_2 and CBM2 have the annotation of dietary fiber as function at destination, while CE17 has no known substrate and CBM48 is involved in glycogen metabolism. Collectively, this points towards increased dietary fiber metabolism in control metagenomes. In CRC metagenomes, CE11, GT19, GH20, GT107, GH123, CBM40 and GH33 were among the most consistently enriched CAZymes. We could not assign a function at destination to any of these CAZyme families, apart from GH20, which is involved in both dietary fiber and peptidoglycan metabolism.

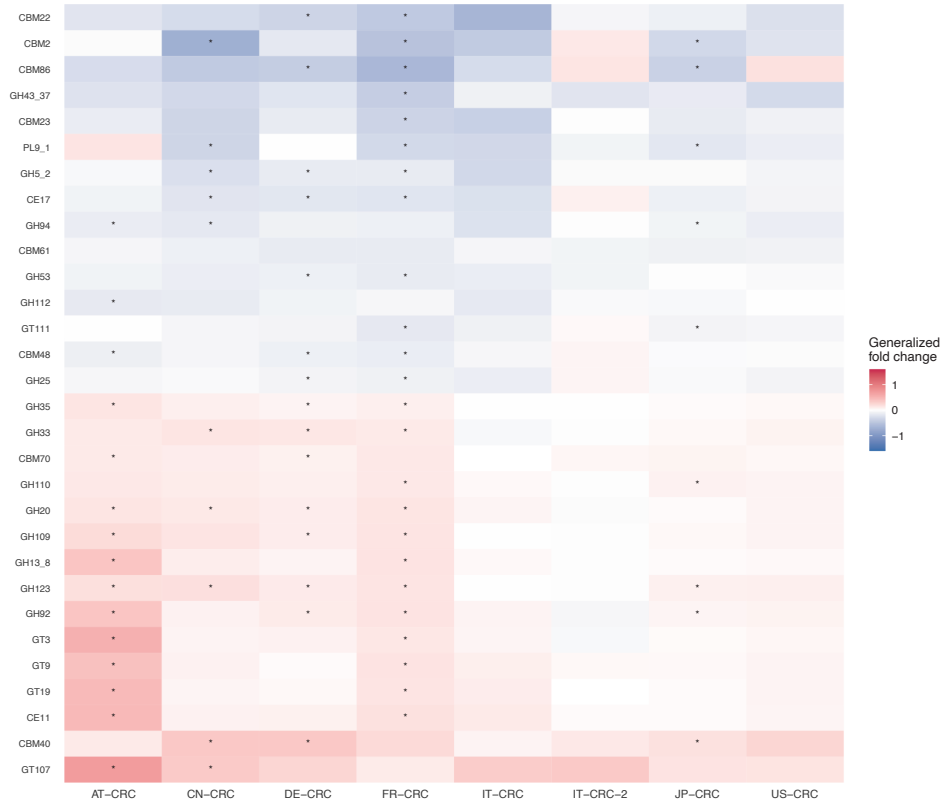


**Figure 2:** Meta-analysis on a per-study basis using Wilcoxon tests and FDR correction for p-values. We selected CAZymes to be displayed by taking the 15 most significantly differentially abundant CAZymes in both CRC metagenomes and in control metagenomes as obtained by the blocked Wilcoxon test using combined data from all eight cohorts. Note that both Italian cohorts contained significantly differentially abundant CAZymes, but these were not among the 15 most significant ones in either direction. The different studies are indicated in the columns of the heatmap and labeled underneath (AT; Austria, CN; China, DE; Germany, FR; France, IT; Italy, JP; Japan and US; United States). Red indicates CAZymes that are more abundant in CRC patients and blue indicates CAZymes more abundant in controls. CAZymes are ordered on the y-axis based on average generalized fold change values.

## Enrichment analysis reveals decreased dietary fiber metabolism and increased glycosaminoglycan (GAG) metabolism in CRC patients

We next investigated whether dietary fiber, and other substrates, were enriched at the substrate level rather than the individual CAZyme level. To this end, we performed enrichment testing through GSEA. Leveraging our suggestion for novel manual substrate annotations in combination with GSEA, we could investigate enrichment of CAZymes at the substrate category level rather than at the individual CAZyme level. Enrichment analysis was performed at the level of the glycan's function at

destination, with testing performed for five large substrate categories (dietary fiber, glycosaminoglycans, peptidoglycan, glycogen and mucin). The difference in number of CAZymes annotated with these specific substrates is large, with 266 unique CAZymes being involved in metabolism of dietary fiber, 27 in glycosaminoglycan, 25 in glycogen, 16 in peptidoglycan and two in mucin. GSEA testing revealed an enrichment of dietary fiber metabolism in controls (adjusted p-value 7.44e^-04), while GAG and peptidoglycan were significantly enriched in CRC patients (adjusted p-value 3.68^-04 and 3.52e^-02, respectively) (Figure 3). GAG are important components of the extracellular matrix of animal tissues (for example in connective and skeletomuscular tissue) and, importantly, have never been detected in plants[33, 34].

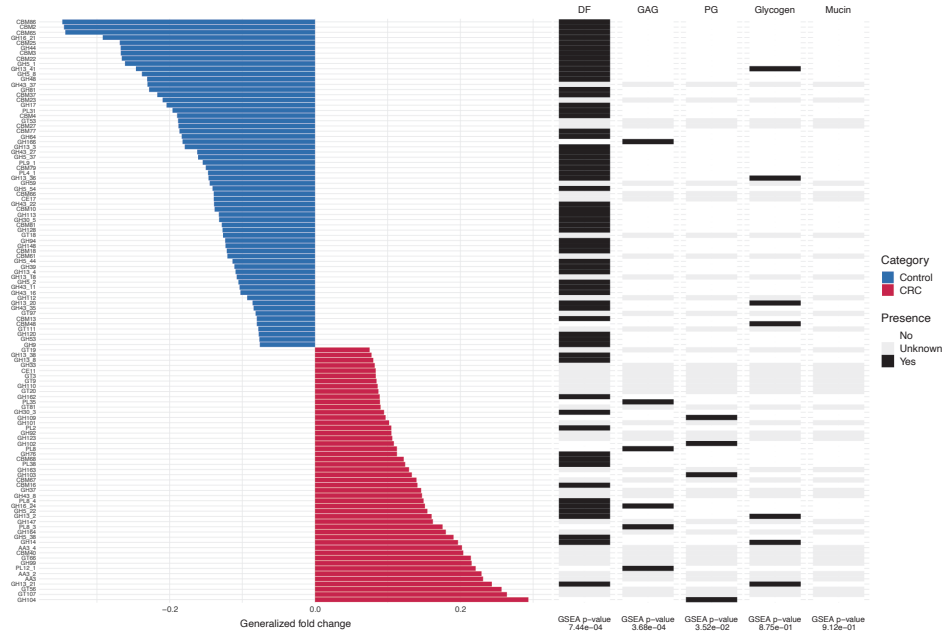


**Figure 3:** Differentially abundant CAZymes and substrate enrichment testing. Horizontal bars show significant abundance differences in CAZymes depicting the effect size by the generalized fold change, as identified by a blocked Wilcoxon test and FDR correction (left) and results of GSEA analysis per substrate category (right) with their respective indication whether a substrate was annotated by a specific CAZyme or not. GSEA was performed on all CAZymes which had a substrate annotation and not only the ones displayed in this plot. CAZymes with 'Unknown' substrate annotation were not included in GSEA. Dietary fiber (DF) was significantly enriched in controls, while glycosaminoglycans (GAG) and peptidoglycan (PG) were significantly enriched in CRC patients. On the right side of the plot, black indicates a CAZyme to be involved in specific substrate metabolism, white indicates it is not and grey indicates that no substrate could be assigned to the respective CAZyme. CAZymes on the left side were included if their adjusted p-value was <0.01 and had a generalized fold change of more than 0.075 or lower than -0.075. Red indicates CAZymes that are more abundant in CRC patients and blue indicates CAZymes more abundant in controls.

# Discussion

We present a new tool for profiling CAZymes from shotgun metagenomic data generated from human fecal samples. To interpret these profiles with respect to broader substrate categories, we additionally derived a hierarchical substrate classification scheme and curated substrate information for all CAZyme families found in the human gut based on the scientific literature. To maximize the detection accuracy of CAZymes in uncharacterized (meta-)genomic sequences based on sequence homology, we systematically optimized and validated family-specific HMMs (and their E-value cutoffs) on a large set of manually annotated CAZyme sequences. By leveraging our novel substrate annotation scheme, we were for the first time able to perform enrichment testing with respect to substrate metabolism for CAZymes in metagenomic case-control studies. In an application to eight studies of CRC, these novel tools revealed a clear CRC-associated CAZyme signature that is largely consistent across the geographically diverse studies included in this meta-analysis.



**Advantages of our newly developed tool over dbCAN2**

The most widely used tool for detecting CAZymes in genomic data are dbCAN2 and its predecessor dbCAN[9, 35]. This tool is largely based on building HMMs from CAZy (sub)families from the CAZy database. While it has proven to be extremely valuable for annotating genomic data, the E-value cutoffs have not been optimized for detection accuracy. Here we used blocked cross-validation on the protein sequences from the CAZy database to optimize HMM E-value cutoffs to more accurately identify CAZymes. Going a step further than annotation of genomic data, we devised a profiling framework to enable users to directly profile CAZymes from human gut metagenomic data. This is a major advance, as it allows for straightforward comparative analyses of CAZyme profiles between (groups of) metagenomic samples using statistical tests suitable for metagenomic data. We will ultimately make the workflow presented in this manuscript (from raw read processing to obtaining matrices with CAZyme abundances) publicly available as open-source software to enable other researchers to easily profile the CAZyme repertoire in their metagenome. While we purely focused on the human gut microbiome, we plan for future updates to extend the CAZyme substrate annotations to gene sequences found in metagenomic gene catalogs available for other mammalian gut microbiomes and environmental communities, such as those in the environment (e.g. soil, ocean)[36-41].

**Manual substrate annotations allow for informative grouping at higher functional levels and statistical enrichment testing**

By detecting broader shifts in carbohydrate degradation capabilities of microbial communities, key insights into community function have been obtained, in some cases

with important implications for host physiology[42, 43]. However, this type of analysis is still not very commonly performed on human gut metagenomic data due to the following technical challenges. The combination of a lack of knowledge on substrate usage of specific CAZyme families and grouping CAZymes into families based on protein sequence similarity does not allow for easy interpretation of CAZyme profiles with regard to substrate utilization[7]. Therefore, it would be highly valuable to have a consistent higher-level substrate annotation that can be used to detect trends across families. A common workflow that researchers currently take when trying to assign functional information to CAZymes is to first investigate, through differential abundance testing, which CAZymes are differentially abundant between two groups of interest and then attempt to annotate these CAZymes only[17, 44]. However, this does not leverage a potential 'substrate-level' effect, as while individual CAZymes may not be significantly differentially abundant, their combined effect may be significant at the substrate level[27]. Such enrichment patterns can, however, be detected using our new substrate annotation scheme in combination with Kolmogorov-Smirnov type statistical tests that are also commonly applied in GSEA[27].



**CAZyme and substrate enrichments correspond to and extend on previous nutritional epidemiological studies investigating CRC**

Increased dietary fiber intake has long been linked to a reduced risk of developing CRC, although there is some conflicting evidence from nutritional epidemiological studies[45-48]. However, it has not been systematically investigated whether the gut microbiome may be involved in the physiological metabolic processes underlying this association[17]. Previous research showed specific CAZymes involved in fiber metabolism to be more abundant in controls and several CAZymes involved in host glycan metabolism (e.g. mucin) to be more abundant in CRC patients[17]. We investigated if this could be confirmed by applying our tool on data from eight different studies with patients recruited in seven different countries on three continents. Additionally, we moved away from exclusively testing individual CAZymes, but instead assessed enrichment of specific substrate categories through a GSEA. Importantly, we found dietary fiber metabolism to be increased in controls and an increase in GAG metabolism in CRC metagenomes (Figure 3). The latter possibly reflects increased meat consumption in CRC patients, which has been linked to an increased risk for developing CRC[49-51].

Unfortunately, no food frequency questionnaire or other detailed dietary information was collected from the CRC patients or controls, which prevents us from correlating reported dietary intake to abundance of specific CAZymes or to the observed enrichment in GAG metabolism. Importantly, while we see this enrichment in GAG in CRC metagenomes, it should be noted that this remains an association and cannot be interpreted as a causal relationship. Lastly, we observed an enrichment in peptidoglycan metabolism in CRC

metagenomes, which is in line with a previous observation of an increase in bacterial cell wall components in CRC metagenomes[17], but this result remains difficult to interpret.
In conclusion, our novel tool allows for accurate profiling of CAZymes from metagenomic shotgun sequence data and the annotation scheme enables substrate-based analysis and interpretation. Together, the tool and substrate annotation scheme pave the way for more informative functional analyses from metagenomic data with respect to the CAZyme repertoire. We applied our tool and annotation scheme to investigate the role of CAZymes and substrate metabolism in CRC metagenomes and discovered specific CAZymes and substrates to be enriched and depleted in CRC metagenomes. While both the dietary fiber and GAG signatures are consistent with known dietary risk factors of CRC, it remains unknown whether the gut microbiome mediates the potentially beneficial and harmful effects of these dietary components in CRC development, or whether the gut microbiome simply adapts to the diet and plays no mechanistic role in the diet-CRC relationship. Therefore, future studies should aim to delineate whether the effects of dietary factors on CRC development are microbiome-mediated or not.



# Notes

**Funding**
QD was supported by a FEMS Research and Training Grant. NK and GZ are supported by EMBL core funding, GZ is partially funded by the German Federal Ministry of Education and Research (BMBF; the de. NBI network no. 031A537B and grant no. 031L0181A).

**Author contributions**
QD, NK and GZ conceptualized and designed the study. QD, NK and CS performed all bioinformatic analyses. QD and HT performed substrate annotations and QD wrote the manuscript with contributions from NK and GZ. GZ supervised the study. All authors discussed and approved the manuscript.

# Supplementary tables

**Table S1**: Table with annotations at several functional categories per glycan. Dietary fiber (DF), non-starch polysaccharides (NSP), peptidoglycan (PG), glycosaminoglycan (GAG).

| Glycan | Origin | Function_in_ origin | Function_ destination_1 | Function_d estination_2 | Function_ destination_3 |
|---|---|---|---|---|---|
| arabinogalactan | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin |
| arabinoxylan | Plant | Structural | DF | NSP | Hemicellulose |
| beta-glucan | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose |
| beta-mannan | Plant | Structural | DF | NSP | Mannans_and_ Heteromannans |
| cellobiose | Plant | Structural | DF | NSP | Cellulose |
| cellulose | Plant | Structural | DF | NSP | Cellulose |
| galactomannan | Plant,Fungal | Structural | DF | NSP | Mannans_and_ Heteromannans |
| glucomannan | Plant,Fungal,Bacteria | Structural | DF | NSP | Mannans_and_ Heteromannans, Hemicellulose |
| xylan | Plant | Structural | DF | NSP | Hemicellulose |
| xyloglucan | Plant | Structural | DF | NSP | Hemicellulose |
| amylopectin | Plant | Storage | DF | Resistant_starch | Unknown |
| amylose | Plant | Storage | DF | Resistant_starch | Unknown |
| arabinan | Plant | Structural | DF | NSP | Pectin |
| glycogen | Animal | Storage | Glycogen | Unknown | Unknown |
| pullulan | Fungal | Structural | DF | Unknown | Unknown |
| starch | Plant | Storage | DF | Resistant_starch | Unknown |
| alginate | Algae,Bacteria | Structural | DF | NSP | Gum |
| carrageenan | Algae | Structural | DF | NSP | Gum |
| galactan | Algae,Plant | Structural | DF | NSP | Pectin |
| laminarin | Algae | Storage | DF | NSP | Unknown |
| porphyran | Algae | Structural | DF | NSP | Unknown |
| fucoidan | Algae | Structural | DF | NSP | Unknown |
| ulvan | Algae | Structural | DF | NSP | Unknown |
| alpha-mannan | Fungal | Structural | DF | NSP | Unknown |
| chitin | Animal,Fungal | Structural | DF | Unknown | Unknown |
| chitosan | Animal,Fungal | Structural | DF | Unknown | Unknown |
| chitodextrin | Animal,Fungal | Structural | DF | Unknown | Unknown |
| dextrin | Plant | Storage | DF | Resistant_ oligosaccharides | Unknown |
| chondroitin | Animal | GAG | GAG | Unknown | Unknown |
| chondroitin_sulfate | Animal | GAG | GAG | Unknown | Unknown |
| heparin | Animal | GAG | GAG | Unknown | Unknown |
| heparin_sulfate | Animal | GAG | GAG | Unknown | Unknown |
| hyaluronan | Animal,Bacteria | GAG | GAG | Unknown | Unknown |
| mucin | Animal | Unknown | Mucin | Unknown | Unknown |
| fructan | Plant | Storage | DF | NSP | Unknown |
| inulin | Plant | Storage | DF | Resistant_ oligosaccharides | Unknown |
| homogalacturonan | Plant | Structural | DF | NSP | Pectin |
| rhamnogalacturonan | Plant | Structural | DF | NSP | Pectin |
| melibiose | Fungal | Unknown | Unknown | Unknown | Unknown |

| Glycan | Origin | Function_in_ origin | Function_ destination_1 | Function_d estination_2 | Function_ destination_3 |
|---|---|---|---|---|---|
| raffinose | Plant | Unknown | DF | Resistant_ oligosaccharides | Unknown |
| agarose | Algae | Structural | DF | NSP | Unknown |
| peptidoglycan | Microbial | Structural | PG | PG | PG |
| lichenin | Plant | Storage | DF | NSP | Unknown |
| GlcNAc | Animal,Bacteria | Structural | Unknown | Unknown | Unknown |
| GalNAc | Animal,Bacteria | Structural | Unknown | Unknown | Unknown |
| sialic_acid | Animal,Bacteria, Archaea,Algae | Unknown | Unknown | Unknown | Unknown |

**Table S2**: Table with annotations for all CAZymes annotated in the IGC (when substrates could be assigned).

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_1 | All_domains_of_life | Unknown | Unknown | Unknown | Unknown | broad |
| GH_2 | All_domains_of_life | Unknown | Unknown | Unknown | Unknown | broad |
| GH_3 | All_domains_of_life | Unknown | Unknown | Unknown | Unknown | broad |
| GH_4 | All_domains_of_life | Unknown | Unknown | Unknown | Unknown | broad |
| GH_5 | All_domains_of_life | Structural, Storage | DF,Glycogen | NSP | Hemicellulose,Cellulose, Unknown | cellulose |
| GH_5_1 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin |
| GH_5_2 | Plant,Bacteria,Fungal, Animal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin,chitosan |
| GH_5_4 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin,xylan,xyloglucan |
| GH_5_5 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin |
| GH_5_7 | Plant,Fungal,Bacteria | Structural | DF | NSP | Mannans_and_ Heteromannans | beta-mannan,mannan,galactomannan,glucomannan |
| GH_5_8 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Mannans_ and_ Heteromannans,Unknown | beta-glucan,cellulose,lichenin,mannan,galactomannan, glucomannan |
| GH_5_13 | Unknown | Unknown | Unknown | Unknown | Unknown | alpha-arabinoside,beta-galactofuranoside |
| GH_5_18 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-mannoside |
| GH_5_21 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_5_22 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin,xylan |
| GH_5_23 | | | | | | glucopyranoside |
| GH_5_25 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Mannans_ and_ Heteromannans,Unknown | beta-glucan,cellulose,lichenin,mannan,galactomannan, glucomannan |
| GH_5_26 | Plant,Microbial, Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin |
| GH_5_35 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_5_36 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose,Mannans_ and_ Heteromannans | beta-glucan,mannan,galactomannan,glucomannan |
| GH_5_37 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_5_38 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin |
| GH_5_41 | Plant,Fungal,Bacterial | Structural | DF | NSP | Mannans_and_ Heteromannans | mannan,galactomannan,glucomannan |
| GH_5_44 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| GH_5_45 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | beta-glucoside,alpha-arabinofuranoside,arabins,arabinoxylan,arabinogalactan |
| GH_5_46 | Plant,Bacteria, Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | beta-glucan,cellulose,lichenin |
| GH_5_52 | Plant,Bacteria, Fungal | Structural | DF | NSP | Hemicellulose,Cellulose | beta-glucan,cellulose |
| GH_5_54 | Plant | Structural | DF | NSP | Cellulose | cellulose |
| GH_8 | Plant,Bacteria, Fungal, Animal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | cellulose,beta-glucan,xylan,lichenin,chitosan |
| GH_9 | Plant,Bacterial, Fungal | Structural | DF | NSP | Hemicellulose,Cellulose, Mannans_and_ Heteromannans | cellulose,xylan,beta-glucan,xyloglucan,glucomannan |
| GH_10 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_11 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_12 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | cellulose,lichenin,beta-glucan,xyloglucan |
| GH_13 | All_domains_of_life | Storage, Structural | DF,Glycogen | Unknown | Unknown | broad |
| GH_13_2 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_polysaccharides |
| GH_13_3 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| GH_13_4 | Plant | | DF | Unknown,Resistant starch | | amylosucrose |
| GH_13_5 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_polysaccharides |
| GH_13_8 | Plant | Storage | DF | Resistant_starch | Unknown | amylose |
| GH_13_9 | Plant | Storage | DF | Resistant_starch | Unknown | amylose |
| GH_13_10 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_ polysaccharides,trehalose |
| GH_13_11 | Plant,Animal | Storage | DF,Glycogen | Resistant_starch, Unknown | Unknown | glycogen,amylopectin |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_13_12 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | starch,glycogen,related_ polysaccharides,pullulan, amylopectin |
| GH_13_13 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | pullulan,amylopectin,glycogen |
| GH_13_14 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | pullulan,amylopectin,glycogen,starch, related_polysaccharides |
| GH_13_16 | | | | | | maltose |
| GH_13_18 | | | | | | sucrose,glucosylglycerate |
| GH_13_19 | Plant | Storage | DF | Resistant_starch | Unknown | amylose,amylaceous_polysaccharides |
| GH_13_20 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | pullulan,amylopectin,glycogen,starch, related_polysaccharides,maltodextrin |
| GH_13_21 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_polysaccharides |
| GH_13_26 | | | | | | alpha-glucan |
| GH_13_27 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_polysaccharides |
| GH_13_28 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | pullulan,amylopectin,glycogen,starch, related_polysaccharides |
| GH_13_29 | Plant | Storage | DF | Resistant_starch | Unknown | trehalose,starch |
| GH_13_30 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| GH_13_31 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_ polysaccharides,sucrose |
| GH_13_32 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | starch,glycogen,related_ polysaccharides,pullulan, amylopectin,amylaceous_polysaccharides |
| GH_13_33 | | | | | | maltose |
| GH_13_36 | Animal,Fungal,Plant | Storage, Structural | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_ polysaccharides,pullulan, maltodextrin |
| GH_13_37 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_polysaccharides |
| GH_13_38 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| GH_13_39 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | pullulan,amylopectin,glycogen,starch, related_polysaccharides |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_13_41 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | pullulan,amylopectin,glycogen,starch, related_polysaccharides |
| GH_13_42 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_ polysaccharides, amlyaceous_polysaccharides |
| GH_14 | Animal,Plant | Storage | Glycogen,DF | Unknown,Resistant_ starch | Unknown | starch,glycogen,related_ polysaccharides |
| GH_15 | Plant | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | trehalose,alpha-glucan,dextran,isomaltos accharide |
| GH_16_3 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Unknown | lichenin,beta-glucan,lamirin,paramylon,pa chyman,keratan_sulfate |
| GH_16_6 | Algae,Plant | Structural | DF | NSP | Pectin | transglycosylases,glucan,galactan |
| GH_16_8 | Algae,Plant | Structural | DF | NSP | Pectin | transglycosylases,glucan,galactan |
| GH_16_12 | Algae | Structural | DF | NSP | Unknown | porphyran |
| GH_16_14 | Algae | Structural | DF | NSP | Unknown | agarose |
| GH_16_15 | Algae | Structural | DF | NSP | Unknown | agarose |
| GH_16_16 | Algae | Structural | DF | NSP | Unknown | agarose,porphyran |
| GH_16_17 | Algae | Structural | DF | NSP | Gum | carrageenan |
| GH_16_21 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Unknown | lichenin,beta-glucan,lamirin,xylan |
| GH_16_24 | Plant,Bacteria, Fungal, Algae,Animal,Bacteria | Structural, GAG | DF,GAG | NSP,Unknown | Hemicellulose,Gum, Unknown | beta-glucan,karatan,hyaluronan,carrage enan |
| GH_16_25 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Unknown | lichenin,beta-glucan |
| GH_17 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Unknown | beta-glucan,lichenin,abscisate |
| GH_18 | Animal,Fungal, Microbial | Structural | DF,PG | PG | Unknown,PG | chitin,chitodextrin,peptidoglycan,high-mannose_glycoproteins |
| GH_19 | Animal,Fungal, Microbial | Structural | DF,PG | PG | Unknown,PG | chitin,chitodextrin,peptidoglycan |
| GH_20 | Bacterial,Animal | Structural | DF,PG | PG | Unknown,PG | gangliosides,chitobiose |
| GH_23 | Animal,Fungal, Microbial | Structural | DF,PG | PG | Unknown,PG | chitin,chitodextrin,peptidoglycan |
| GH_24 | Animal,Fungal, Microbial | Structural | DF,PG | PG | Unknown,PG | chitodextrin,peptidoglycan |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_25 | Animal,Fungal, Microbial | Structural | DF,PG | PG | Unknown,PG | chitodextrin,peptidoglycan |
| GH_26 | Plant,Fungal,Bacteria | Structural | DF | NSP | Mannans_and_ Heteromannans, Hemicellulose | mannan,galactomannan,glucomannan,beta-mannan,xylan |
| GH_27 | Animal,Plant,Unknown | Structural | DF | NSP | Gum,Pectin,Hemicellulose,Unknown | alpha-galactoside,N-acetylgalactosamine,alpha-glucan,arabinopyranoside |
| GH_28 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan,pectin, homogalacturonan, polygalacturote |
| GH_29 | Algae,Animal | Structural | DF | NSP | Unknown, HMO | fucoidan |
| GH_30_1 | Plant,Algae | Structural | DF | NSP | Hemicellulose,Pectin | xylan,glucan,galactan,glucuronide |
| GH_30_2 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_30_3 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| GH_30_4 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-fucoside |
| GH_30_5 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin | arabinogalactan,beta-galactan |
| GH_30_6 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-glucoside |
| GH_30_8 | Plant | Structural | DF | NSP | Hemicellulose | xylan,glucuronoarabinoxylans |
| GH_30_9 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin | beta-glucuronoside |
| GH_31 | Plant | Storage | DF | Resistant_starch | Unknown | starch,xylose,sucrose,isomaltose |
| GH_32 | Plant | Storage | DF | Resistant_ oligosaccharides | Unknown | sucrose,inulin,levan |
| GH_33 | Animal,Microbial, Algae | Unknown | Unknown | Unknown | HMO,Unknown | sialic_acid |
| GH_34 | Animal,Microbial, Algae | Unknown | Unknown | Unknown | Unknown | sialic_acid |
| GH_35 | Animal,Fungal | Structural | DF | | Pectin, Unknown | beta-galactoside,alpha-arabinoside,chitosan |
| GH_36 | Plant | Unknown | DF | NSP, Resistant_ oligosaccharides | Gum, Unknown | alpha-galactoside,N-acetylgalactoseamine,raffinose |
| GH_37 | Animal,Bacteria,Plant, Fungal | Storage | Unknown | Unknown | Unknown | trehalose |
| GH_38 | Fungal | Structural | DF | NSP | Unknown | alpha-mannan |
| GH_39 | Plant | Structural | DF | NSP | Hemicellulose | xylan,dermatan_sulfate |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_42 | Plant,Animal | Unknown | Unknown | Unknown | Unknown | beta-galactoside,alpha-arabinoside |
| GH_43 | Plant,Bacteria | Structural | DF | NSP | Hemicellulose,Pectin,Gum | broad |
| GH_43_1 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside,arabin, arabinoxylan,arabinogalactan |
| GH_43_2 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_3 | Unknown | Unknown | Unknown | Unknown | Unknown | galactofuranoside |
| GH_43_4 | Unknown | Unknown | Unknown | Unknown | Unknown | arabin |
| GH_43_5 | Unknown | Unknown | Unknown | Unknown | Unknown | arabin |
| GH_43_7 | Plant | Structural | DF | NSP | Hemicellulose | xylan,arabinofuranoside |
| GH_43_8 | | | | | | galactofuranoside |
| GH_43_9 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_10 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside,arabin, arabinoxylan,arabinogalactan |
| GH_43_11 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside,arabin, arabinoxylan,arabinogalactan |
| GH_43_12 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside,arabin, arabinoxylan,arabinogalactan |
| GH_43_16 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan, xylobiose,arabinofuranoside, arabin, arabinoxylan,arabinogalactan |
| GH_43_17 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_18 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_19 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_22 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside, arabin,arabinoxylan,arabinogalactan |
| GH_43_23 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_24 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin | arabinogalactan |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_43_26 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_27 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside, arabin,arabinoxylan,arabinogalactan |
| GH_43_28 | Plant | Structural | DF | NSP | Hemicellulose | xylan,arabinofuranoside |
| GH_43_29 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside, arabin,arabinoxylan,arabinogalactan |
| GH_43_30 | Unknown | Unknown | Unknown | Unknown | Unknown | galactofuranoside |
| GH_43_31 | Unknown | Unknown | Unknown | Unknown | Unknown | galactofuranoside |
| GH_43_33 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_43_34 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinofuranoside,arabin,arabinoxylan, arabinogalactan,galactofuranoside |
| GH_43_35 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xylan,xylobiose,arabinofuranoside, arabin,arabinoxylan,arabinogalactan |
| GH_43_37 | Unknown | Unknown | Unknown | Unknown | Unknown | galactofuranoside,arabin |
| GH_44 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | cellulose,lichenin,beta-glucan, xyloglucan |
| GH_45 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Mannans_ and_ Heteromannans,Unknown | cellulose,lichenin,beta-glucan, xyloglucan,mannan,galactomannan, glucomannan |
| GH_46 | Animal,Fungal | Structural | DF | Unknown | Unknown | chitosan |
| GH_47 | Fungal | Structural | DF | NSP | Unknown | alpha-mannan |
| GH_48 | Plant,Bacteria, Fungal, Animal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Unknown | cellulose,lichenin,beta-glucan,chitin, chitodextrin |
| GH_49 | Fungal | Structural | DF | Unknown | Unknown | dextran,pullulan |
| GH_50 | Algae | Structural | DF | NSP | Unknown | agarose |
| GH_51 | Plant,Bacteria,F ungal | Structural, Storage | DF | NSP | Gum,Pectin,Hemicellulose,Cellulose, Unknown | cellulose,lichenin,beta-glucan,xylan, arabinofuranoside,arabin,arabinoxylan, arabinogalactan |
| GH_52 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_53 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin | arabinogalactan |
| GH_54 | Plant | Structural | DF | NSP | Hemicellulose | xylan,alpha-arabinosides |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_55 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan,lamirin,paramylon, pachyman |
| GH_57 | Plant | Storage | DF | Resistant_starch | Unknown | amylose,galactoside |
| GH_58 | Animal,Bacteria, Archaea,Algae | Unknown | Unknown | Unknown | Unknown | sialic_acid |
| GH_59 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-galactosides,ceramide_derivatives |
| GH_63 | Plant | Storage | DF | Resistant_starch | Unknown | starch,alpha-glucan,mannosylglycerate, glucosylglycerate |
| GH_64 | Unknown | Unknown | DF | NSP | Hemicellulose | lamirin,paramylon,pachyman |
| GH_65 | Unknown | Unknown | Unknown | Unknown | Unknown | matlose,trehalose,kojibiose |
| GH_66 | Bacteria | Unknown | DF | Unknown | Unknown | alpha-glucan,dextran |
| GH_67 | Plant | Structural | DF | NSP | Hemicellulose | alpha-glucuronoside,xylan |
| GH_68 | Unknown | Unknown | DF | Unknown | Unknown | sucrose,beta-fructofuranoside |
| GH_70 | Unknown | Unknown | DF | Unknown | Unknown | sucrose,alpha-glucan |
| GH_71 | Unknown | Unknown | Unknown | Unknown | Unknown | isolichenin,pseudonigeran,nigeran |
| GH_73 | Animal,Fungal, Microbial | Structural | DF,PG | PG | Unknown,PG | peptidoglycan,chitodextrin,high-mannose_ glycopeptides |
| GH_74 | Plant,Bacteria,Fungal | Structural, Storage | DF | NSP | Hemicellulose,Cellulose, Mannans_ and_ Heteromannans,Unknown | cellulose,lichenin,beta-glucan, xyloglucan,mannan,galactomannan, glucomannan |
| GH_76 | Plant,Fungal | Storage Structural | DF | Resistant_starch,NSP | Unknown | alpha-mannan,starch |
| GH_77 | Unknown | Unknown | Unknown | Unknown | Unknown | alpha-glucan |
| GH_78 | Plant | Structural | DF | NSP | Pectin | alpha-rhamnoside,rhamnogalacturonan |
| GH_79 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | beta-glucuronoside,heparin_sulfate, hyaluronate |
| GH_80 | Animal,Fungal | Structural | DF | | Unknown | chitosan |
| GH_81 | Unknown | Unknown | DF | NSP | Hemicellulose | lamarin,paramylon,pachyman |
| GH_82 | Algae | Structural | DF | NSP | Gum | carrageenan |
| GH_83 | Animal,Bacteria, Archaea,Algae | Unknown | Unknown | Unknown | Unknown | sialic_acid |
| GH_84 | Animal,Bacteria | Unknown | Unknown | Unknown | Unknown | glycoproteins |
| GH_85 | Plant,Unknown | Unknown | Unknown | Unknown | Unknown | high-mannose_glycopeptides |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_86 | Algae | Structural | DF | NSP | Unknown | agarose,porphyran |
| GH_87 | Unknown | Unknown | Unknown | Unknown | Unknown | alpha-glucan,isolichenin, pseudonigeran,nigeran |
| GH_88 | Unknown | GAG | GAG | Unknown | Unknown | glycosaminoglycans |
| GH_89 | Animal | Unknown | Mucin | Unknown | Unknown | mucin |
| GH_90 | Unknown | Unknown | Unknown | Unknown | Unknown | rhamnose |
| GH_91 | Plant | Storage | DF | Resistant_ oligosaccharides | Unknown | inulin |
| GH_92 | Unknown | Unknown | Unknown | Unknown | Unknown | glycoproteins |
| GH_93 | Unknown | Unknown | Unknown | Unknown | Unknown | alpha-arabin |
| GH_94 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose,Cellulose | cellobiose,cellodextrin,chitobiose, beta-glucan |
| GH_95 | Plant,Algae,Animal | Structural | DF | NSP | Hemicellulose,Unknown, HMO | fucoidan,fucoside,xyloglucan,HMO, blood_group_glycoconjugates |
| GH_96 | Algae | Structural | DF | NSP | Unknown | agarose |
| GH_97 | Unknown | Unknown | Unknown | Unknown | Unknown | maltose |
| GH_98 | Plant | Structural | DF | NSP | Hemicellulose | beta-galactoside,blood_group_antigen, xylan |
| GH_99 | Unknown | Unknown | Unknown | Unknown | Unknown | glucose-substituted_mannose,fungal_ cell_wall |
| GH_101 | Unknown | Unknown | Unknown | Unknown | Unknown | N-acetylgalactoseamine |
| GH_102 | Microbial | Structural | PG | PG | PG | peptidoglycan |
| GH_103 | Microbial | Structural | PG | PG | PG | peptidoglycan |
| GH_104 | Microbial | Structural | PG | PG | PG | peptidoglycan |
| GH_105 | Plant,Bacteria,Algae | Structural | DF | NSP | Gum,Pectin,Unknown, Pectin | rhamnogalacturonan,ulvan, arabinogalactan |
| GH_106 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| GH_108 | Animal,Fungal, Microbial | Structural | DF,PG | ,PG | Unknown,PG | peptidoglycan,chitodextrin |
| GH_109 | Microbial | Structural | PG | PG | PG | peptidoglycan |
| GH_110 | Unknown | Unknown | Unknown | Unknown | Unknown | blood_group_antigen |
| GH_111 | Unknown | Unknown | Unknown | Unknown | Unknown | N-acetylglucoseamine |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_112 | Unknown | Unknown | Unknown | Unknown | Unknown | galacto-N-biose,lacto-N-biose |
| GH_113 | Plant,Fungal,Bacteria | Structural | DF | NSP | Mannans_and_ Heteromannans | mannan,galactomannan, glucomannan |
| GH_115 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_116 | Unknown | Unknown | Unknown | Unknown | Unknown | Unknown |
| GH_117 | Algae | Structural | DF | NSP | Gum,Unknown | beta-galactofuranoside,agarose, carrageenan |
| GH_120 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| GH_121 | Plant | | DF | | | hydroxproline-rich_glycoproteins |
| GH_122 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| GH_123 | Unknown | Unknown | Unknown | Unknown | Unknown | N-acetylgalactoseamine |
| GH_125 | Fungal | Structural | DF | NSP | Unknown | alpha-mannan |
| GH_126 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| GH_127 | Plant | Unknown | DF | NSP | Pectin, Hemicellulose | Unknown |
| GH_128 | Unknown | Unknown | DF | NSP | Hemicellulose | lamirin,paramylon,pachyman |
| GH_129 | Animal,Microbial | Unknown, Structural | Mucin,PG | Unknown,PG | Unknown,PG | mucin,peptidoglycan |
| GH_130 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-mannosides |
| GH_133 | Animal | Storage | Glycogen | Unknown | Unknown | glycogen |
| GH_136 | Animal,Unknown | Unknown | Unknown | Unknown | HMO,Unknown | lacto-N-tetraose |
| GH_137 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| GH_138 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| GH_139 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| GH_140 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| GH_141 | Plant,Algae | Structural | DF | NSP | Hemicellulose,Unknown | xylan,fucoidan |
| GH_142 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| GH_143 | Unknown | Unknown | Unknown | Unknown | Unknown | rhamnogalacturonan |
| GH_144 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| GH_145 | Unknown | Unknown | DF | NSP | Gum,Unknown | gum-arabic |
| GH_146 | Unknown | Unknown | Unknown | Unknown | Unknown | sugar_beet_arabin |
| GH_147 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-galactoside |
| GH_148 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| GH_149 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan,lamirin |
| GH_150 | Algae | Structural | DF | NSP | Gum | carrageenan |
| GH_151 | Unknown | Unknown | Unknown | Unknown | HMO,Unknown | blood_group_antigen |
| GH_153 | Unknown | Unknown | Unknown | Unknown | Unknown | Unknown |
| GH_154 | Plant,bacteria | structural | DF | NSP | Gum, pectin | beta-glucuronoside |
| GH_156 | Animal,Bacteria, Archaea,Algae | Unknown | Unknown | Unknown | Unknown | sialic_acid |
| GH_157 | Unknown | Unknown | DF | NSP | Hemicellulose | lamirin,paramylon,pachyman |
| GH_158 | Unknown | Unknown | DF | NSP | Hemicellulose | lamirin,paramylon,pachyman |
| GH_159 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-galactofuranosides |
| GH_160 | Bacteria | Unknown | Unknown | Unknown | Unknown | bacterial_polysaccharides |
| GH_161 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan,lamirin |
| GH_162 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| GH_163 | Unknown | Unknown | Unknown | Unknown | Unknown | Unknown |
| GH_164 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-mannosides |
| GH_165 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-galactoside |
| GH_166 | Bacteria | GAG | GAG | Unknown | Unknown | bacterial_polysaccharides |
| GH_167 | Algae | Structural | DF | NSP | Gum | carrageenan |
| PL_1_2 | Plant | Structural | DF | NSP | Pectin | pectate |
| PL_1_6 | Plant | Structural | DF | NSP | Pectin | pectate |
| PL_2 | Plant | Structural | DF | NSP | Pectin | pectate |
| PL_2_1 | Plant | Structural | DF | NSP | Pectin | pectate |
| PL_2_2 | Plant | Structural | DF | NSP | Pectin | pectin |
| PL_4 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| PL_4_1 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| PL_4_4 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| PL_6_1 | Algae,Bacteria | Structural | DF | NSP | Gum | dermatan_sulfate,alginate |
| PL_6_2 | Unknown | Unknown | Unknown | Unknown | Unknown | Unknown |
| PL_8 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | hyaluron,chondroitin,xanthan, chondroitin_sulfate,dermatan_ sulfate,heparin,heparin_sulfate |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| PL_8_1 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | hyaluron,chrondoitin |
| PL_8_2 | Animal | GAG | GAG | Unknown | Unknown | heparin,chrondoitin,dermatan |
| PL_8_3 | Animal | GAG | GAG | Unknown | Unknown | chondroitin_sulfate,dermatan_sulfate |
| PL_8_4 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate |
| PL_9_1 | Plant | Structural | DF | NSP | Pectin | pectate,rhamnogalacturonan |
| PL_9_2 | Plant | Structural | DF | NSP | Pectin | Unknown |
| PL_9_4 | Plant | Structural | DF | NSP | Pectin | Unknown |
| PL_10 | Plant | Structural | DF | NSP | Pectin | pectate |
| PL_10_1 | Plant | Structural | DF | NSP | Pectin | pectate |
| PL_10_2 | Plant | Structural | DF | NSP | Pectin | pectate |
| PL_11 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| PL_11_1 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| PL_11_2 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |
| PL_12 | Animal | GAG | GAG | Unknown | Unknown | heparin,heparin_sulfate |
| PL_12_1 | Animal | GAG | GAG | Unknown | Unknown | heparin |
| PL_12_2 | Animal | GAG | GAG | Unknown | Unknown | heparin,heparin_sulfate |
| PL_13 | Animal | GAG | GAG | Unknown | Unknown | heparin,heparin_sulfate |
| PL_15 | Algae,Bacteria,Animal | Structural,GAG | DF,GAG | NSP,Unknown | Gum,Unknown | heparin,heparin_sulfate,alginate |
| PL_15_1 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate |
| PL_15_2 | Animal | GAG | GAG | Unknown | Unknown | heparin,heparin_sulfate |
| PL_16 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | hyaluron,chondroitin |
| PL_17 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate |
| PL_17_1 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate |
| PL_17_2 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate |
| PL_18 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate |
| PL_21 | Animal | GAG | GAG | Unknown | Unknown | heparin,heparin_sulfate |
| PL_21_1 | Animal | GAG | GAG | Unknown | Unknown | heparin,heparin_sulfate |
| PL_22 | Plant | Structural | DF | NSP | Pectin | pectin |
| PL_22_1 | Plant | Structural | DF | NSP | Pectin | pectin |
| PL_23 | Animal | GAG | GAG | Unknown | Unknown | chondroitin |
| PL_26 | Plant | Structural | DF | NSP | Pectin | rhamnogalacturonan |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| PL_27 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin | arabinogalactan |
| PL_28 | Algae | Structural | DF | NSP | Unknown | ulvan |
| PL_29 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | chondroitin_sulfate,dermatan_sulfate, hyaluronic_acid |
| PL_30 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | hyaluron,chondroitin |
| PL_31 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate,beta-glucuron |
| PL_32 | Algae,Bacteria | Structural | DF | NSP | Gum | alginate |
| PL_33 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | hyaluron,gellan,chondroitin_sulfate |
| PL_33_1 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | hyaluron,chondroitin |
| PL_33_2 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | chondroitin_sulfate,dermatan_sulfate, hyaluronan |
| PL_35 | Animal | GAG | GAG | Unknown | Unknown | chondroitin |
| PL_37 | Algae,Animal | Structural,GAG | DF,GAG | NSP,Unknown | Unknown | heparan_sulfate,dermatan_sulfate, chondroitin_sulfate, ulvan |
| PL_38 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin | glucuronan |
| PL_40 | Algae | Structural | DF | NSP | Unknown | ulvan |
| CE_1 | Unknown | Unknown | Unknown | Unknown | Unknown | Unknown |
| CE_2 | Plant | Structural | DF | NSP | Hemicellulose | xylan,xylooligosaccharides |
| CE_3 | Plant | Structural | DF | NSP | Hemicellulose | xylan,xylooligosaccharides |
| CE_4 | Plant,Animal,Fungal, Microbial | Structural | DF,PG | NSP,PG | Hemicellulose,Unknown,PG | peptidoglycan,chitin,xylan |
| CE_5 | Plant | Structural | DF | NSP | Hemicellulose | xylan,xylooligosaccharides,cutin |
| CE_6 | Plant | Structural | DF | NSP | Hemicellulose | xylan,xylooligosaccharides |
| CE_7 | Plant | Structural | DF | NSP | Hemicellulose | xylan,xylooligosaccharides, cephalosporin |
| CE_8 | Plant | Structural | DF | NSP | Pectin | pectin |
| CE_9 | Microbial | Structural | PG | PG | PG | peptidoglycan |
| CE_11 | Bacterial | Unknown | Unknown | Unknown | Unknown | LPS |
| CE_12 | Plant | Structural | DF | NSP | Hemicellulose,Pectin | pectin,rhamnogalacturonan,xylan, xylooligosaccharides |
| CE_14 | Unknown | Unknown | Unknown | Unknown | Unknown | Unknown |
| CE_15 | Plant,Bacteria | structural | DF | NSP | Gum, pectin | lignin,lignocellulose |
| CE_17 | Unknown | Unknown | Unknown | Unknown | Unknown | Unknown |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| CBM_2 | Plant,Animal,Fungal | Structural | DF | NSP | Cellulose,Hemicellulose, Unknown | cellulose,chitin,xylan |
| CBM_3 | Plant | Structural | DF | NSP | Cellulose | cellulose |
| CBM_4 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | xylan,beta-glucan |
| CBM_5 | Animal,Fungal | Structural | DF | Unknown | Unknown | chitin |
| CBM_6 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | xylan,beta-glucan |
| CBM_9 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| CBM_10 | Plant | Structural | DF | NSP | Cellulose | cellulose |
| CBM_12 | Animal,Fungal | Structural | DF | Unknown | Unknown | chitin |
| CBM_13 | Plant | Structural | DF | NSP | Hemicellulose | mannose,xylan |
| CBM_15 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| CBM_16 | Plant,Fungal,Bacteria | Structural | DF | NSP | Cellulose,Mannans_and_ Heteromannans | cellulose,glucomannan |
| CBM_18 | Animal,Fungal | Structural | DF | Unknown | Unknown | chitin |
| CBM_20 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_21 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_22 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | xylan,beta-glucan |
| CBM_23 | Unknown | Unknown | Unknown | Unknown | Unknown | mannan |
| CBM_25 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_26 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_27 | Unknown | Unknown | Unknown | Unknown | Unknown | mannan |
| CBM_30 | Plant | Structural | DF | NSP | Cellulose | cellulose |
| CBM_32 | Plant | Structural | DF | NSP | Mannans_and_ Heteromannans | galactose,lactose,lamirin,pustulan, beta-mannan,mannooligosaccharides |
| CBM_34 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_35 | Plant | Structural | DF | NSP | Hemicellulose | xylan,manns,mannooligosaccharides, beta-galactan |
| CBM_36 | Plant | Structural | DF | NSP | Hemicellulose | xylan,xylooligosaccharides |
| CBM_37 | Plant,Animal,Fungal | Structural | DF | NSP, | Cellulose,Hemicellulose, Unknown | xylan,chitin,cellulose,alfalfa_cell_walls, ba_stems,wheat_straw |
| CBM_38 | Plant | Storage | DF | Resistant_ oligosaccharides | Unknown | inulin |
| CBM_40 | Animal,Bacteria, Archaea,Algae | Unknown | Unknown | Unknown | Unknown | sialic_acid |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| CBM_41 | Plant,Animal,Fungal | Storage, Structural | DF,Glycogen | Resistant_starch, Unknown | Unknown | alpha-glucan,starch,glycogen, amylose,pullulan |
| CBM_42 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | arabinoxylan,arabin,arabinogalactan |
| CBM_47 | Algae | Structural | DF | NSP | Unknown | fucoidan |
| CBM_48 | Animal | Storage | Glycogen | Unknown | Unknown | glycogen |
| CBM_50 | Animal,Fungal, Microbial | Structural | DF,PG | PG | Unknown,PG | peptidoglycan,chitin |
| CBM_51 | Unknown | Unknown | Unknown | Unknown | Unknown | galactose,blood_group_antigen |
| CBM_52 | Plant,Bacteria ,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| CBM_54 | Plant,Animal, Fungal | Structural | DF | NSP, | Hemicellulose,Unknown | xylan,chitin,yeast_cell_wall |
| CBM_55 | Animal,Fungal | Structural | DF | Unknown | Unknown | chitin |
| CBM_56 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| CBM_57 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-glucuronide |
| CBM_58 | Unknown | Unknown | Unknown | Unknown | Unknown | maltoheptaose,acarbose |
| CBM_61 | Unknown | Unknown | Unknown | Unknown | Unknown | beta-galactan |
| CBM_62 | Plant,Bacteria | Structural | DF | NSP | Gum,Pectin,Hemicellulose | xyloglucan,arabinogalactan,galactomann |
| CBM_63 | Plant | Structural | DF | NSP | Hemicellulose,Cellulose | cellulose,arabinoxylan |
| CBM_64 | Plant | Structural | DF | NSP | Cellulose | cellulose |
| CBM_65 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | xyloglucan,beta-glucan |
| CBM_66 | Unknown | Unknown | Unknown | Unknown | Unknown | inulan,levan |
| CBM_67 | Unknown | Unknown | Unknown | Unknown | Unknown | rhamnose,mannose |
| CBM_68 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_69 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_70 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | hyaluron |
| CBM_71 | Unknown | Unknown | Unknown | Unknown | Unknown | lactose,LacNAc |
| CBM_72 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose,Cellulose | cellulose,beta-glucan,xylan,beta-mannan |
| CBM_73 | Animal,Fungal | Structural | DF | Unknown | Unknown | chitin |
| CBM_75 | Plant | Structural | DF | NSP | Hemicellulose | xyloglucan |
| CBM_76 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| CBM_77 | Plant | Structural | DF | NSP | Pectin | pectin |
| CBM_79 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |

| (Sub) family | Origin_Glycan | Function_in_ origin | Function_ destination_1 | Function_ destination_2 | Function_destination_3 | Glycan_annotation |
|---|---|---|---|---|---|---|
| CBM_80 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| CBM_81 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| CBM_82 | Plant | Storage | DF | Resistant_starch | Unknown | starch |
| CBM_84 | Unknown | Unknown | Unknown | Unknown | Unknown | xanthan |
| CBM_85 | Plant,Bacteria,Fungal | Structural | DF | NSP | Hemicellulose | beta-glucan |
| CBM_86 | Plant | Structural | DF | NSP | Hemicellulose | xylan |
| CBM_87 | Animal,Bacteria | GAG | GAG | Unknown | Unknown | galactoseaminogalactan |

# References


1. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nat Rev Microbiol. 2013;11(7):497-504.
2. Makki K, Deehan EC, Walter J, Bäckhed F. The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease. Cell Host Microbe. 2018;23(6):705-15.
3. Ríos-Covián D, Ruas-Madiedo P, Margolles A, Gueimonde M, de Los Reyes-Gavilán CG, Salazar N. Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health. Front Microbiol. 2016;7:185.
4. Litvak Y, Byndloss MX, Bäumler AJ. Colonocyte metabolism shapes the gut microbiota. Science. 2018;362(6418).
5. Morrison DJ, Preston T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes. 2016;7(3):189-200.
6. Desai MS, Seekatz AM, Koropatkin NM, Kamada N, Hickey CA, Wolter M, et al. A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. Cell. 2016;167(5):1339-53.e21.
7. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42(Database issue):D490-5.
8. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J. 1991;280 ( Pt 2)(Pt 2):309-16.
9. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46(W1):W95-w101.
10. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80.
11. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026-8.
12. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32(8):834-41.
13. Coelho LP, Alves R, Monteiro P, Huerta-Cepas J, Freitas AT, Bork P. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. Microbiome. 2019;7(1):84.
14. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;preprint arXiv:1303.3997.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
16. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
17. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10(11):766.
18. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, et al. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PLoS One. 2016;11(5):e0155362.
19. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat Commun. 2015;6:6528.
20. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. 2017;66(1):70-8.
21. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25(6):968-76.
22. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019;25(4):667-78.
23. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25(4):679-89.
24. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. A Lego System for Conditional Inference. The American Statistician. 2006;60(3):257-63.
25. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. Microbiome meta-analysis and cross-disease

comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol. 2021;22(1):93.
26. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B (Methodological). 1995;57(1):289-300.
27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.
28. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. BioRxiv. 2021:060012.
29. Ausland C, Zheng J, Yi H, Yang B, Li T, Feng X, et al. dbCAN-PUL: a database of experimentally characterized CAZyme gene clusters and their substrates. Nucleic Acids Res. 2021;49(D1):D523-d8.
30. Koropatkin NM, Cameron EA, Martens EC. How glycan metabolism shapes the human gut microbiota. Nat Rev Microbiol. 2012;10(5):323-35.
31. Patova OA, Golovchenko VV, Ovodov YS. Pectic polysaccharides: structure and properties. Russ Chem Bull. 2015;63(9):1901-24.
32. Stephen AM, Champ MM, Cloran SJ, Fleith M, van Lieshout L, Mejborn H, et al. Dietary fibre in Europe: current state of knowledge on definitions, sources, recommendations, intakes and relationships to health. Nutr Res Rev. 2017;30(2):149-90.
33. Yamada S, Sugahara K, Ozbek S. Evolution of glycosaminoglycans: Comparative biochemical study. Commun Integr Biol. 2011;4(2):150-8.
34. Namburi RB, Berteau O, Spillmann D, Rossi M. Chondroitinase AC: A host-associated genetic feature of Helicobacter bizzozeronii. Vet Microbiol. 2016;186:21-7.
35. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012;40(Web Server issue):W445-51.
36. Lesker TR, Durairaj AC, Gálvez EJC, Lagkouvardos I, Baines JF, Clavel T, et al. An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome. Cell Rep. 2020;30(9):2909-22.e6.
37. Li J, Zhong H, Ramayo-Caldas Y, Terrapon N, Lombard V, Potocki-Veronese G, et al. A catalog of microbial genes from the bovine rumen unveils a specialized and diverse biomass-degrading environment. Gigascience. 2020;9(6).
38. Pan H, Guo R, Zhu J, Wang Q, Ju Y, Xie Y, et al. A gene catalogue of the Sprague-Dawley rat gut metagenome. Gigascience. 2018;7(5).
39. Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, Dauser S, Schuster N, Walzer C, et al. Large-Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity. mSystems. 2020;5(6).
40. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science. 2015;348(6237):1261359.
41. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol. 2021;4(1):604.
42. von Schwartzenberg RJ, Bisanz JE, Lyalina S, Spanogiannopoulos P, Ang QY, Cai J, et al. Caloric restriction disrupts the microbiota and colonization resistance. Nature. 2021;595(7866):272-7.
43. Wastyk HC, Fragiadakis GK, Perelman D, Dahan D, Merrill BD, Yu FB, et al. Gut-microbiota-targeted diets modulate human immune status. Cell. 2021.
44. Bhattacharya T, Ghosh TS, Mande SS. Global Profiling of Carbohydrate Active Enzymes in Human Gut Microbiome. PLoS One. 2015;10(11):e0142038.
45. Kunzmann AT, Coleman HG, Huang WY, Kitahara CM, Cantwell MM, Berndt SI. Dietary fiber intake and risk of colorectal cancer and incident and recurrent adenoma in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. Am J Clin Nutr. 2015;102(4):881-90.
46. Aune D, Chan DS, Lau R, Vieira R, Greenwood DC, Kampman E, et al. Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. BMJ. 2011;343:d6617.
47. Ben Q, Sun Y, Chai R, Qian A, Xu B, Yuan Y. Dietary fiber intake reduces risk for colorectal adenoma: a meta-analysis. Gastroenterology. 2014;146(3):689-99.e6.
48. Goodlad RA. Dietary fibre and the risk of colorectal cancer. Gut. 2001;48(5):587-9.
49. Chao A, Thun MJ, Connell CJ, McCullough ML, Jacobs EJ, Flanders WD, et al. Meat consumption and risk of colorectal cancer. JAMA. 2005;293(2):172-82.
50. Sandhu MS, White IR, McPherson K. Systematic review of the prospective cohort studies on meat consumption

and colorectal cancer risk: a meta-analytical approach. Cancer Epidemiol Biomarkers Prev. 2001;10(5):439-46.
51. Gurjao C, Zhong R, Haruki K, Li YY, Spurr LF, Lee-Six H, et al. Discovery and features of an alkylating signature in colorectal cancer. Cancer Discov. 2021.