Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Interrogating gut bacterial genomes for discovery of novel carbohydrate degrading enzymes

Ana S Luis and Eric C Martens

Individual human gut bacteria often encode hundreds of enzymes for degrading different polysaccharides. Identification of co-localized and co-regulated genes in these bacteria has been a successful approach to identify enzymes that participate in full or partial saccharification of complex carbohydrates, often unmasking novel catalytic activities. Here, we review recent studies that have led to the discovery of new activities from gut bacteria and summarize a general scheme for identifying gut bacteria with novel catalytic abilities, locating the enzymes involved and investigating their activities in detail. The strength of this approach is amplified by the availability of abundant genomic and metagenomic data for the human gut microbiome, which facilitates comparative approaches to mine existing data for new or orthologous enzymes.

Address
University of Michigan Medical School, Ann Arbor, MI 48109, USA

Corresponding author: Martens, Eric C (emartens@umich.edu)

Current Opinion in Chemical Biology 2018, **47C**:126–133

This review comes from a themed issue on **Mechanistic Biology**

Edited by **John A. Gerlt** and **Tobias** J. **Erb**

https://doi.org/10.1016/j.cbpa.2018.09.012

Introduction

Symbiotic bacteria inhabiting the distal guts of humans and other mammals are responsible for the majority of non-starch dietary fiber polysaccharide digestion [1,2]. As such, the genomes of individual gut bacteria often harbor many dozens or hundreds of individual carbohydrate-active enzymes (CAZymes) to carry out this catalysis [3•,4]. While the general structural features of plant cell wall and other fiber polysaccharides have been well characterized, the range of enzymatic activities that degrade them, especially very complex structures [5••] or botanical source-specific variants [6,7], are still being defined. In this review, we highlight how detailed investigations of individual human gut bacteria have provided a useful path towards discovery of new CAZyme activities. This emerging knowledge is critical to fully understand the range and evolution of these catalytic abilities and also provides fundamental insight into human digestion.

Carbohydrate active enzyme classification

Degradation of the array of polysaccharides available in the human gut is possible through the combination of various CAZymes with different specificities. These enzymes fall into two main groups, glycoside hydrolases (GHs) and polysaccharide lyases (PLs), although carbohydrate esterases (CEs) also play important roles on some substrates. GHs catalyze the hydrolysis of glycosidic bonds by either a single or a double-displacement mechanism leading to inversion or retention of the stereochemistry of the anomeric carbon, respectively [8]. PLs are non-hydrolytic enzymes that cleave uronic acid-containing polysacharides via a β-elimination mechanism yielding an unsaturated uronic acid at the non-reducing end [9,10]. CAZymes are classified into sequence-based families via the continuously updated CAZy database [11]. The members of a family display the same fold, same catalytic apparatus and mechanism [12,13]. One of the few exceptions is GH97 that contains enzymes with two different mechanisms [14]. GH families displaying the same fold, mechanism and catalytic residues but a low sequence similarity are grouped into clans [11,13]. Additionally, enzyme topology is related to endo-modes or exo-modes of action. Endo-acting enzymes cleave internal linkages within the substrate and have an open cleft; whereas, exo-acting hydrolases cleave at end of the chain and present an active site pocket that only accommodates the terminal sugar [15].

A 'functional microbiology' approach to discovering new carbohydrate modifying enzymes

Whole genome transcriptional profiling techniques, so called 'functional genomics', are now routine since the advent of microarray and subsequent next-generation sequencing based methods (e.g. RNAseq). These techniques allow researchers to quickly locate genes that are activated under different physiological conditions, which for symbiotic human gut bacteria often involves large changes in gene expression in response to polysaccharide cues [3•,16••,17,18,19••]. Using this approach, gene clusters that encode CAZymes and other functions related to polysaccharide degradation have been identified in a number of gut bacteria, spanning most common dietary polysaccharides [1,20•] and even more exotic structures such as seaweed-derived agarose and porphyran [21,22••]. Many studies have been conducted in gut Bacteroidetes,

a prominent saccharolytic phylum in the human colon, in which individual members typically group all the genes required for the recognition, uptake and degradation of a specific glycan into co-regulated polysaccharide utilization loci (PULs), gene clusters that encode, in addition to CAZymes, TonB-dependent transporters, SusD family binding proteins and other functions [17]. However, additional studies are also emerging in members of the abundant Firmicutes, Actinobacteria and Verrucomicrobia phyla, revealing different types of gene arrangements and encoded enzymatic machinery such as enzyme-associated ABC-transport systems and cellulosome-like systems [23,24•,25,26].

Since many complex dietary polysaccharides contain several different sugars and glycosidic linkages—therefore requiring multiple degradative enzymes—identifying transcriptionally active gene clusters with multiple CAZymes often reveals suites of enzymes that work together and often uncovers genes encoding enzymes with novel catalytic roles (Figure 1). In addition, similar polysaccharides sometimes vary in subtle ways between botanical sources; identifying a gene cluster associated with degradation of one form of a polysaccharide enables searching of other genomic or metagenomic data to find orthologous, but partially variant gene clusters that may have adapted to other forms of the same polysaccharide. An example of this is utilization of the dicotyledonous hemicellulose xyloglucan. Growth, transcriptional profiling and molecular genetics with a single strain of *Bacteroides ovatus* (strain ATCC8483) revealed a 12-gene PUL (xyloglucan utilization locus; XyGUL), containing 8 different CAZymes, that is necessary for growth on a simplified form of xyloglucan from tamarind seed [19••,27••]. While the originally identified *B. ovatus* PUL encoded enzymes for removing α-L-arabinofuranosyl side-chain linkages present in solanaceous plants (tomatoes, eggplants), it did not encode enzymes for removal of α-fucosyl linkages attached to the xyloglucans present in leafy greens like lettuce [28]. However, comparative genomics using the originally discovered *B. ovatus* XyGUL revealed the presence of homologous and often highly syntenic PULs in other sequenced gut bacteria that also encode enzymes in known α-fucosidase families (glycoside hydrolase families 29 and 95), suggesting adaptation to the modifications specific to lettuce. Moreover, the increasing availability of high-quality metagenomic sequencing data from human fecal samples, like those from the Human Microbiome Project, enables even deeper discovery of novel locus variants without the need for direct cultivation [27••]. Taken together, the scheme shown in Figure 1 illustrates a path to connect the ability of pure or mixed enrichment cultures to grow on complex polysaccharides with the genes involved and then further compare across genomes and metagenomes to discover variant loci. For several very complex plant cell wall polysaccharides discussed below, this process has led to discovery of new catalytic activities that have only been uncovered in human gut bacteria in the past few years.

Plant cell wall xylan: new activities inside existing families

During growth on wheat xylan, a major component of monocotyledonous plant cell walls that exhibits structural variation between plant sources [29,30], *B. ovatus* upregulates two PULs [19••] that encode 19 different predicted enzymes, although one was shown to be catalytically inactive (Figure 1). In addition to known xylanolytic activities, two of the enzymes required for xylan utilization revealed new activities inside previously characterized GH families (Figures 1–2). A member of family GH98, previously known to only include blood group A- or B-cleaving *endo*-β-D-galactosidases [31], was shown to generate long xylo-oligosaccharides from corn arabinoxylan, requiring the D-xylose (D-Xyl) of the leaving group to be doubly linked to both α1,2-L-arabinofuranose (L-Ara*f*) and α1,3-D-Xyl [7]. A novel GH95 enzyme was found to cleave L-gal). Previously, this family only contained α-L-fucosidases [32]. The catalytic apparatus is fully conserved inside members of this family. Indeed, a single difference in the −1 subsite, a Thr in the α-L-galactosidase that is replaced by a His in α-L-fucosidases, is speculated to be the substrate specificity determinant inside family GH95 (Figure 2a) [7].

Novel families of rhamnosidases

Several plant cell wall polysaccharides that are common in the human diet contain α-L-rhamnose (L-Rha) and a number of *Bacteroides thetaiotaomicron* PULs containing candidate rhamnosidases have been previously discovered [19••]. Cleavage of rhamnosidic linkages is enzymatically challenging because the axial C2-hydroxyl causes a destabilizing *syn*-diaxial orientation in the transition state [33]. Until recently only GH78, 90 and GH106 families have been shown to contain α-L-rhamnosidases [34,35,36]. The discovery of two new families active on L-Rha-α1,4-D-glucuronic acid (D-GlcA) moieties that cap arabinogalactan polysaccharide side chains revealed novel rhamnosidase features (Table 1, Figure 2b). Members of PL27, that are unrelated to any of the previously known PL or GH families, overcome the challenge of cleaving L-Rha through the β-elimination mechanism. Briefly, the catalytic base (Tyr) abstracts the proton at C5 of D-GlcA, generating a double bond between C4 and C5 (Δ4,5-anhydroglucuronic acid) that leads to the elimination of the L-Rha at C4 [37]. Additionally, the recently described GH145 family is proposed to use His as the only catalytic residue (nucleophile) and the carboxylate of the leaving group (D-GlcA) fulfills the catalytic acid/base role. A completely novel feature of these enzymes is the active site location at the posterior surface of the seven bladed β-propeller and not in the highly conserved anterior pocket, suggesting that these enzymes might have an additional ancestral activity that is not yet known [38].