Genome-centric Portrait of the Microbes' Cellulolytic Competency

Yubo Wang

Westlake University

Liguan Li

Technical University of Denmark

Yu Xia

Southern University of Science and Technology

Feng Ju

Westlake University

Tong Zhang (✉ zhangt@hku.hk)

Environmental Microbiome Engineering and Biotechnology Laboratory, Centre for Environmental Engineering Research, University of Hong Kong Shenzhen Institute of Research and Innovation, Shenzhen, China

Research article

Keywords: Cellulolytic, synergy mechanisms, genome-centric, function interpretation

Posted Date: October 7th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-81485/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Neither the abundance of the exo/endoglucanase GH modules nor the taxonomy affiliation is informative enough in inferring whether a genome is of a potential cellulolytic microbe or not. By interpreting the complete genomes of 2642 microbe strains whose phenotypes have been well documented, we are trying to reveal a more reliable genotype and phenotype correlation on the specific function niche of cellulose hydrolysis.

Results: By incorporating into the annotation approach an automatic recognition of the potential synergy machineries, a more reliable prediction on the corresponding microbes' cellulolytic competency could be achieved. The potential cellulose hydrolyzing microbes could be categorized into 5 groups according to the varying synergy machineries among the carbohydrate active modules/genes annotated. Results of the analysis on the 2642 genomes revealed that some cellulosome gene clusters were in lack of the surface layer homology module (SLH) and microbe strains annotated with such cellulosome gene clusters were not certainly cellulolytic. Hypothesized in this study was that cellulosome-independent genes harboring both the SLH module (mediate the attachment of the enzymes to the host microbe's cell surface) and the cellulose-binding carbohydrate binding module (mediate the attachment of the enzymes to the cellulose substrate) were likely an alternative gene apparatus initiating the formation of the cellulose-enzyme-microbe (CEM) complexes; and their role is important especially for the cellulolytic anaerobes without cellulosome gene clusters.

Conclusions: In the genome-centric prediction on the corresponding microbes' cellulolytic activity, recognition of the synergy machineries that include but are not limited to the cellulosome gene clusters is equally important as the annotation of the individual carbohydrate active modules or genes. This is the first time that a pipeline was developed for an automatic recognition of the synergy among the carbohydrate active units annotated. With promising resolution and reliability, this pipeline should be a good add to the bioinformatic tools for the genome-centric interpretations on the specific function niche of cellulose hydrolysis.

Background

In the era of high-throughput sequencing, the genetic information that is inherently whispering hints of the microbes' function niches is becoming easily accessible [1, 2]. However, the bottleneck remains largely on properly identifying and characterizing these genetic hints and inferring the microbes' function potentials. In this study, we focus on the genome-centric interpretation on the specific function niche of cellulose hydrolysis. Traditional approaches, including the microscope observation, cultivation of the cellulose-degrading microbes, as well as purification and characterization of the cellulolytic enzymes [3, 4], have set a good foundation in understanding how the microbes and their enzymes may interact with the cellulosic substrates. Although it is believed that most of the cellulolytic microbes may still be hiding in plain sight due to the isolation bottleneck, access to their genome information has opened a new window to shed light on them.

Regarding to the genome-centric interpretation on the function niche of cellulose hydrolysis, current annotation approaches focus on tapping the diversity and the abundance of the individual carbohydrate active enzyme (CAZy) modules annotated. Applying the HMMsearch-based dbCAN annotation platform, referring to the well-curated CAZy database [5, 6], a decent amount of information on the abundances of the diverse CAZy modules in a genome could be obtained. And a more recent method for the annotation and interpretation of CAZymes, named as CUPP, has also been developed [7]. However, often encountered in practice was a lack of confidence in predicting the microbes' real cellulolytic competency based solely on the abundances of the relevant CAZy modules annotated. For example, a total number of 21 exo/endoglucanase GH modules in the genome of *Actinoplanes missouriensis* 431 could not point to a conclusion that this strain was able to hydrolyze cellulose [8]; and it remains a puzzle why *Clostridium acetobutylicum*, with the cellulosome gene cluster identified in its genome, do not have the cellulose degrading capability [9–11].

What is in lack in current genome-centric interpretations is the recognition of the synergy among the individual CAZy modules and the synergy among the carbohydrate active genes harboring these CAZy modules, although synergism is one of the highly-appreciated features in efficient cellulose hydrolysis [12, 13]. Cellulolytic enzymes are known as modular proteins, the most straightforward synergy would occur among the diverse CAZy modules in one single gene/enzyme; e.g., if one gene has both the cellulose-binding module CBM6 and the exoglucanase module GH48, the CBM6 could help bring this GH48 to its action site. Based on their co-occurrence patterns across genomes and metagenome datasets, Sebastian et al. achieved targeted discovery of functional modules of plant biomass-degrading protein families [14]. A higher level of synergy would occur among the diverse carbohydrate active enzymes in one microbe, on which aspect, cellulosome is the most highly recognized synergy machinery in anaerobes [15]; and the CEM complex initiated by hypha penetration is the more commonly observed synergy mechanism in aerobic cellulolytic *Fungi* [12]. Although the carbohydrate active enzymes did not assemble into one entity as those in the cellulosome complexes [10, 16], in the CEM complex of some aerobic *Fungi*, physical closeness among the individual carbohydrate active enzymes sandwiched in between the Fungus cells and the cellulose substrates makes the synergy among these enzymes possible [15, 17–21].

May the formation of the cellulosome-independent CEM complexes be possible in anaerobes? This question is raised in the context of the fact that the number of the cellulolytic anaerobes is much larger than the number of anaerobes annotated with cellulosome gene clusters. It is not common to observe in anaerobes the physical apparatus like hypha to facilitate the physical penetration as in Fungus, having been reported in literature was that the hypothesized glycocalyx mediated microbe-cellulose contact in anaerobes [22]. Most cellulolytic species were of their optimal growth rates when they adhere to the cellulosic substrate, and the microbe-cellulose contact is important for the host microbes to get easy access to the enzymatic hydrolyzing products [23, 24]. It has also been reported that the excreted free cellulase would contribute little to the microbes' cellulolytic activity [12, 23]. One of the objectives of this study is, by investigating complete genomes of the 2642 microbe strains whose phenotypes have been well characterized, to propose potential alternative genetic machineries (in lieu of the cellulosome

complexes) that may initiate the CEM complex formation through the microbe-cellulose adhesion, especially in anaerobes.

The cell-surface-anchoring role of the SLH module has been proposed and investigated, which suggests that the SLH module could mediate the attachment of the enzymes to the host microbe's cell surface [25, 26]. Among the various carbohydrate binding modules (CBMs), a subset of the CBM modules are cellulose-binding, herein, referred to as cCBM in this paper, and they could mediate the attachment of the enzymes to the cellulose substrate. Our hypothesis is that enzymes harboring both the SLH module and the cCBM (cellulose-binding CBM module) should be able to facilitate the formation of the CEM complexes: through the SLH module, the enzyme is anchored to the host microbe's cell surface, and through its cellulose-binding cCBM module, the enzyme and the anchored microbe cell could be directed to its cellulose substrate; in this way, the microbe-cellulose adhesion would be mediated by such enzymes; and all other carbohydrate active enzymes excreted by the host microbe would be sandwiched in the confined area between the microbe cell and its cellulose substrate, forming the CEM complexes. Considering that physical closeness in the form of the CEM complex is critical for microbial cellulose hydrolysis, the role of such SLH-cCBM genes, as one of the potential alternative machineries for the CEM complex formation, in predicting the corresponding microbes' cellulolytic competency might be underestimated in current genome-centric interpretation on the function niche of cellulose hydrolysis.

Physical link or physical closeness is important for the synergy interactions among the carbohydrate active units [27]. One recent progress in the recognition of the physical-link among the carbohydrate active genes was the establishment of the polysaccharide-utilization loci (PUL) database [28]. In this study, we are trying to introduce the annotation of two more features regarding the physical connections among the CAZy modules and among the carbohydrate active enzymes: 1) clustering patterns among the CAZy modules along genes; and 2) machineries that may facilitate the assembly or physical aggregation of the diverse carbohydrate active enzymes in one microbe.

Starting with the analysis of the complete genomes of the 2642 microbe strains, we are aiming to test the possibility of developing an annotation pipeline for: 1) an automatic recognition of the clustering patterns among the CAZy modules in carbohydrate active genes in genomes, 2) recognition of potential alternative genetic machineries, in lieu of cellulosomes, for the CEM complex formation in microbes, and 3) recognition and categorization of the genomes of potential cellulolytic microbes. The applicability of the pipeline in the annotation of metagenome assembled genomes (MAGs) would be further tested with the annotation of 7904 reference genomes downloaded from NCBI.

Results

Co-occurring patterns among the CAZy modules in the carbohydrate active genes

Genes are the basic units encoding enzymes, presented in Fig. 1 and the Appendix file 1 are the frequencies at which CAZy modules co-occurring with each other in same genes; and these frequencies were calculated from the CAZy modules in carbohydrate active genes annotated in the 2642 complete genomes. One of the most distinctive co-occurrences was observed between the exo/endo-glucanase GH modules (GH6, GH9 and GH48) and the cellulose binding cCBMs modules (dominantly CBM2, CBM3 and CBM30). Among the CAZy modules annotated in the 2642 complete genomes, 51% of the GH48 modules were observed being present in same genes with the CBM2 module; and CBM2 was also observed in 29% of the genes harboring the GH6 module. This is in accord with the reported importance of the cCBM modules in: 1) the initiation of the exo/endoglucanase GH modules' hydrolytic activity and 2) the progressiveness of the exoglucanase along the cellulose chains [12]. Similarly observed was the co-occurrence between the xylanase GH modules (e.g., GH53, GH10) and the hemicellulose binding CBM modules (e.g., CBM61, CBM22), e.g., 26% of the genes harboring the GH10 module would also carry the CBM22 module.

Besides their high frequencies co-occurring with the cellulose-binding cCBM modules, GH9 and GH48 were also the two modules with the highest frequencies co-occurring with the dockerin module, e.g. ~23% of the GH9-harboring genes were also identified with the dockerin module; and this suggested that GH9 and GH48 might be the two most common catalytic components in the cellulosome complexes. Collaboration between the exoglucanase and the endoglucanase was another important synergy pattern in cellulose hydrolysis; and this corresponded with the observation that ~ 20% of the exoglucanase GH48 module coexisted in same genes as the endoglucanase GH74 module.

Categorization of the carbohydrate active genes

Part of the visualization of the CAZy module arrangement along carbohydrate active genes is demonstrated in Fig. 2. According to the CAZy modules they harbor, the carbohydrate active genes could be classified into two broad categories: genes of the cellulosome gene clusters and carbohydrate active genes independent of the cellulosome gene clusters. As is summarized in Table 1, the cellulosome gene clusters consist of two parts, the scaffold genes (A1, A2 and A3) and genes of the catalytic components (A-s: dockerin + GH/cCBM). The scaffold genes in the cellulosome gene clusters could be further categorized into three types ('A1', 'A2' and 'A3'), according to whether the SLH module is initially in (type 'A1') or at least could be incorporated (type 'A2') into these scaffold genes. The integration of the 'A2-a' gene and the 'A2-b' gene by the dockerin and cohesion modules would incorporate the SLH module into the type 'A2' scaffolds. The scaffold genes of type 'A3' are in lack of the SLH module.

Table 1
Categorization of genes based on the CAZy modules they harbor

Gene type			Abundance of modules in one gene					Gene Description
			SLH	Dokerin	Cohesin	cGH	cCBM	
Cellulosomal gene clusters	Scaffold gene	A1	>=1	0	>=3	>=0	>=0	Adhering scaffold
								(Cohesin + Cohesin+... +Cohesin + SLH)
		A2-a	>=0	>=1	>=3	>=0	>=0	Adhering scaffold
								(Cohesin + Cohesin+... +Cohesin + Dockerin)
		A2-b	>=1	>=0	>=1	>=0	>=0	Counterparts of A2-a
								(Cohesin + SLH)
	Catalytic constituents	A3	0	0	>=3	>=0	>=0	Free scaffold
								(Cohesin + Cohesin+... +Cohesin)
		A-s	>=0	>=1	>=0	>=1	>=0	Catalytic components to be assembled
								(Dockerin + GH)
Gene sets independent of cellulosomes	B	>=1	>=0	>=0	>=0	>=1	microbe-cellulose adhesion	
								(SLH + cCBM)
		C	>=0	>=0	>=0	>=1	>=1	Cellulose binding cellulase
							(GH + cCBM)	

Note: The cCBM referred to in this table are only the cellulose-binding CBM module, and the cGH module in this table indicate only the exo/endoglucanase GH modules.

	D	>=0	>=0	>=0	>=1	0	Free cellulase (Solitude GH)
Note: The cCBM referred to in this table are only the cellulose-binding CBM module, and the cGH module in this table indicate only the exo/endoglucanase GH modules.							

Among the carbohydrate active genes independent of the cellulosome gene clusters, what might have been underestimated was the role of genes (type ‘B’) harboring both the SLH module and the cellulose-binding cCBM modules. Theoretically, enzymes encoded by these SLH-cCBM genes could adhere onto the microbes’ cell surface through its SLH module, and the cellulose-binding cCBM counterpart could help drag the SLH-attached microbe cell to its cellulosic substrates. Such microbe-cellulose adhesion facilitated by these SLH-cCBM enzymes might help sandwich the excreted carbohydrate enzymes in between the microbe cell and the cellulosic substrate, in which way the CEM complex would form. It is reasonable to speculate that, similar as the hypha mediated CEM complex, the SLH-cCBM mediated CEM complex may provide the same physical closeness needed for the synergy among enzymes aggregating in between the microbe cell and the cellulose substrate. There are two other types of cellulosome-independent cellulolytic active genes: type ‘C’ and type ‘D’; both type ‘C’ and type ‘D’ genes harbor the cellulolytic GH modules; and the cellulose-binding cCBM modules were identified in type ‘C’ genes but not in type ‘D’ genes.

Categorization of genomes of potential cellulolytic microbes

As has been summarized in Table S1, among the 2642 microbe strains investigated, only 270 strains were identified with both the exoglucanase GH modules and the endoglucanase GH modules in their genomes. The genomes of these 270 microbe strains harboring both the exoglucanase and endoglucanase GH modules were preliminarily categorized into Group I in this study. Result of the analysis suggested that only genomes in Group I were of potential cellulose hydrolyzing microbes. It was noted that a total number of only one exo/endo GH module (in quite few cases, a total number of two exo/endo GH modules) would be identified in a genome if this genome was annotated with only the exoglucanase GH modules or with only the endoglucanase GH modules, and none of these genomes are of microbe strains with reported cellulolytic activities.

The 270 genomes in Group I could be further categorized into six subgroups (Group I-a, Group I-b, ..., Group I-e), according to the types of carbohydrate active genes they harbor. The criteria for this categorization are summarized in Table 2. Cellulosome gene clusters were identified in genomes of the first three subgroups: Group I-a, Group I-b and Group I-c. Unlike that of the Group I-a genomes in which the scaffold genes were of either type A1 or type A2 (with SLH module), the scaffold-genes in genomes of both Group I-b and Group I-c were of type ‘A3’ (without SLH module). The differentiating feature of genomes in Group I-b and Group I-c is that cellulosome-independent SLH-cCBM genes were identified in

genome of Group I-b, which may act as an alternative microbe-cellulose adhesion machinery; while such cellulosome-independent SLH-cCBM genes were absent in genomes of Group I-c.

Table 2. Sub-categorization of the genomes with both exoglucanase and endoglucanase GH modules.

Abundance of different genes identified					Category	Number of genomes assigned in each group
Adhering scaffold (A1 or A2)	Free scaffold (A3)	SLH-cCBM (B)	cGH-cCBM (C)	Solitude GH (D)		
>=1	>=0	>=0	>=0	>=0	Group I-a	3
0	>=1	>=1	>=1	>=1	Group I-b	2
0	>=1	0	>=1	>=1	Group I-c	4
0	0	>=1	>=1	>=1	Group I-d	16
0	0	>=1	>=1	>=0		
0	0	>=1	>=0	>=1		
0	0	0	>=1	>=1	Group I-e	139
0	0	0	>=1	0		
0	0	0	0	>=1	Group I-f	106

Note: the cellulose-binding CBM module was referred to as cCBM, e.g., the SLH-cCBM genes are genes harboring both the SLH module and the cellulose-binding CBM module; and the ex-/endo-glucanase GH modules were referred to as cGH modules in this table

the other three subgroups (Group I-d, Group I-e and Group I-f) were all free of the cellulosome gene clusters. Among these three subgroups, the SLH-cCBM genes were identified only in genomes of Group I-d; the cellulose binding cCBMs were observed in at least one of the cellulolytic genes in genomes of Group I-e; and genomes of Group I-f were featured with the annotation result that all of their cellulolytic genes were free of the cellulose-binding cCBM modules. A detailed summary on the diversity and abundances of the various carbohydrate active genes annotated, and the categorization of these 270 genomes could be found in the Appendix file 2.

Cellulolytic competency of genomes categorized into the different subgroups

What would the varying genome features indicate on the corresponding microbes' cellulolytic competency? As has been illustrated in the above section, cellosome gene clusters are present in three subgroups: Group I-a, Group I-b and Group I-c. Among the 2642 microbe strains, the three strains assigned to Group I-a: *R. thermocellum* ATCC 27405, *R. thermocellum* DSM 1313 and *C. clariflavum* DSM 19732 were all paradigm cellulolytic microorganisms with tethered cellosome complexes and the highest cellulose hydrolyzing rates reported [29–31]. Both *C. sp.* BNL1100 and *C. cellulolyticum* H10 assigned to Group I-b were reported as proficient cellulose hydrolysers with cellosome complexes observed [32, 33]. There were four strains assigned to Group I-c, being *C. acetobutylicum* ATCC 824, *C. cellulovorans* 743B, *C. acetobutylicum* EA 2018 and *C. acetobutylicum* DSM 1731, respectively; except for *C. cellulovorans* 743B, the three strains of the *C. acetobutylicum* were all inert in crystalline cellulose hydrolysis [9, 32, 34].

Genomes assigned to Group I-d were in two distinct taxonomy groups: strains from the aerobic genus of *Paenibacillus* and strains from the anaerobic genus of *Caldicellulosiruptor*. The seven anaerobic strains in *Caldicellulosiruptor* were all characterized as being cellulolytic [35–41]; and the eight aerobic strains of *Paenibacillus* were principally known as plant growth promoter residing either in soil with rich forest residuals or in plant root systems [10, 42–47]. The mutualism between *Paenibacillus* and the plant may proceed in a way that the bacteria provide growth hormones and antibiotics to plants, and the plant residues may provide the *Paenibacillus* strains with their carbohydrates substrates.

The total number of the exo/endoglucanase GH modules annotated in genomes of Group I-f varied from 2 to 8, and none of their cellulolytic GH modules were in same genes as the cellulose-binding cCBM modules; correspondingly, strains in Group I-f were all inert in cellulose utilization. The total number of the exo/endoglucanase GH modules annotated in genomes of Group I-e were in a range of 2–35, and at least one of its carbohydrate active genes harbored both the cellulolytic GH module and the cellulose-binding cCBM module. The cellulolytic capacity of microbes in Group I-e varied from being non-cellulolytic to polysaccharides-utilizer to cellulolytic. And there was no apparent correlation between the number of the exo/endoglucanase GH modules annotated and the corresponding microbe's cellulolytic capability. For example, *Stercorarium subsp.* DSM8532 was cellulolytic with a total number of only 5 exo/endoglucanase GH modules annotated in its genome [48]; while *Actinoplanes missouriensis* 431 was not able to grow on cellulose although a total number of 21 exo/endoglucanase GH modules were annotated in its genome [48].

Phylogeny of the 2642 genomes were visualized in the circle tree in Fig. 3; genomes assigned into Group I-a, Group I-b, Group I-c, Group I-d and Group I-e were highlighted in different colors; genomes of Group I-f were not highlighted in this circle tree since none of them were cellulolytic. It can be seen from Fig. 3 that the cellulolytic capability was not phylogenetically highly conservative. For example, among the 13 strains in the genus of *Clostridium* (Appendix file 2), 1 of them was assigned to Group I-b, 4 in Group I-c, 1 in Group I-e; and all the other 7 strains in this genus were not cellulolytic. The results further signified that it might not be a workable approach to predict the corresponding microbe's cellulolytic capability based solely on the phylogenetic affiliation of a genome.

The pipeline developed and its application in the annotation of the metagenome assembled genomes on the function niche of cellulose hydrolysis

To facilitate an automatic identification and categorization of the potential cellulolytic genomes, the categorizing criteria proposed in this study were embodied in R scripts. Description of the overall analysis flow and the usage of the scripts could be found in Github. The applicability of this annotation pipeline was further tested with the annotation of the 7904 reference genomes downloaded from NCBI.

When applied to the dbCAN annotation results, this pipeline was very time-efficient in identifying and categorizing genomes of the potential cellulolytic microbes. It took ~ 30 minutes to get: 1) a summary of the diversity and abundances of all the CAZy modules identified in each of these 7902 genomes (Appendix file 3); 2) abundances of the diverse carbohydrate active genes in each genome (Appendix file 3); and 3) assignment of the potential cellulose hydrolyzing genomes into 6 subgroups according to the varying synergy machineries annotated (Appendix file 3). 5 out of these 7902 reference genomes were assigned into Group I-a, 9 genomes were in Group I-b, 15 genomes were in Group I-c, 3 genomes in Group I-d, 15 genomes in Groups I-e and 8 genomes in Group I-f. Figure S1 presents the phylogeny of genomes in the first five categories. Consistent with results of the survey on the 2642 complete genomes, cellulosome-gene clusters were annotated only in a small number of microbes, and the varying cellulolytic capabilities were not phylogenetically conservative[49].

Discussion

Genome-centric features of anaerobes harboring cellulosome gene clusters while being inert in cellulose hydrolysis

Cellulosome complexes by its nature could enable the assembly of a number of carbohydrate active units. In previous genome-centric interpretation on the function niche of cellulose hydrolysis, the presence of the cellulosome gene clusters was always taken as an indicator of the efficient cellulose hydrolysers. However, results of this survey suggested that not all cellulosomal gene clusters and the corresponding cellulosome complexes were of the classical configuration, and a finer classification of the cellulosome gene clusters is needed.

Cellulosomal complexes in lack of the SLH module might not be cell surface adhering, and the formation of the CEM complex should be aided by some cellulosome-independent SLH-cCBM genes as annotated in genome of Group I-b. Free cellulosomal complexes (*C. acetobutylicum* in Group I-c) that could not be held in between the microbe cell and the cellulosic substrate would limit the microbes' access to the enzymatic hydrolyzing products, in which case, the host microbe might become reluctant in the energy-consuming synthesis and assembly of cellulosomes. This may explain why the three *C. acetobutylicum* strains were all inert in cellulose hydrolysis although cellulosome gene clusters were identified in their genomes.

Alternative machinery for the CEM complex formation

The potential role of the cellulosome independent SLH-cCBM genes in initiating the microbe-cellulose contact and the succeeding formation of the CEM complexes was proposed in this study. The cellulosome-independent enzymes encoded by genes (represented by 'SLH-cCBM' gene in this study) harboring both the SLH module and the cellulose-binding cCBM module might be an alternative machinery facilitating the microbe-cellulose adhesion. And such microbe-cellulose contact could further initiate the formation of the CEM complex by sandwiching the carbohydrate active enzymes in between the microbe cells and the cellulosic substrate. The physical closeness in the form of the CEM complex could guarantee: 1) synergy among the enzymes (including the free cellulosome complexes) physically constrained in confined areas, and 2) easy access to the enzymatic hydrolyzing products by the host microbes. And the roles of such SLH-cCBM genes might have been underestimated in current genome-centric interpretation on the function niche of cellulose hydrolysis.

Although rigorous wet-lab test is needed to verify the role of the SLH-cCBM enzymes in initiating the formation of the CEM complexes, the hypothesis proposed in this study could be supported by the annotation that genomes in Group I-b and Group I-d all corresponded with cellulolytic microbes. The cellulosome gene clusters in Group I-b are all free of the SLH module, which means that the corresponding cellulosome is not cell-surface anchoring and could not initiate the CEM complex formation by itself, same as that in genomes of Group I-c. While unlike genomes of Group I-c, several cellulosome-independent SLH-cCBM genes are annotated in genomes of Group I-b, correspondingly, microbes in Group I-b are all cellulolytic while the three *C. acetobutylicum* strains in Group I-c are all inert in cellulose hydrolysis. Similarly, exo/endoglucanase GH modules are annotated in genomes of both Group I-d and Group I-e, the differentiating feature between genomes of Group I-d and Group I-e is not the abundance of the GH/cCBM modules identified, but whether SLH-cCBM genes were annotated along with the exo/endoglucanase GH modules; with the presence of the SLH-cCBM genes, genomes of Group I-d all correspond to cellulytic microbes, while the cellulose-hyrdolyzing capacity of the microbes in Group I-f varies.

The industrial production of acetone/ethanol/proponol by *C. acetobutylicum* strains has been a mature technology, and the possibility of *C. acetobutylicum* being able to ferment cellulose would introduce new possibilities for more sustainable solvent production from cheap substrates that include the agricultural wastes [48]. One scenario proposed in this study to make the *C. acetobutylicum* strains cellulolytic active is by introducing the SLH-cCBM genes into their genomes.

Limitations of the pipeline developed in this study

One limitation of this study is that the annotation in this study is based mainly on the CAZy database and the dbCAN annotation platform, and microbe-cellulose adhesion mechanisms delivered by non-CAZy genes could not be covered by this annotation approach. We do not think there is no other microbe-cellulose adhesion machineries exist except for the SLH-cCBM genes and the cellulosome gene clusters, especially in aerobes like Fungi. However, the current knowledge on these alternative machineries, especially their genetic foundations are limited. For example, glycocalyx containing extracellular

polymeric substances (EPS) was reported as a “glue” between the microbe cell and the cellulosic substrates in *R. albus* 7 [50]. This limitation leads to the uncertainty in the genome-centric interpretation on the cellulolytic capacity of microbes assigned into Group I-c and Group I-e. Novel microbe-cellulose adhesion mechanisms that could not be identified by the approach developed in this study might exist in the cellulolytic microorganisms assigned into Group I-c or Group I-e, e.g., *C. cellulovorans* 743B (Group I-c) and *R. champanellensis* 18P13 (Group I-e). Another factor that needs to be considered in the application of this pipeline is that the quality of the genome matters, more reliable functional interpretation is expected for genomes with higher completeness and lower contamination.

Another limitation is introduced by the promiscuity of the CAZy families, many CAZy families have multiple characterized biochemical functions, such as GH9, which is predominantly endoglucanases (EC 3.2.1.4) but also a few exoglucanases (EC 3.2.1.91), similarly is the family GH5. In this study, as long as the exoglucanase activity of a GH module is listed in CAZy, we would count this module as exoglucanase and did not conduct finer classification of the EC annotation. The consideration behind this approach is that: 1) CAZy is more widely used than EC in the annotation of the carbohydrate active genes and genomes, and the CAZy database is well organized; 2) for the genome-centric interpretation of the corresponding microbes cellulose hydrolyzing capacity, an integration of the diversity and abundance of the CAZy module identified and their potential synergy mechanisms may provide a fair estimation, although definitely there is room for further improvement of the accuracy. This approach could at least be taken as a trade-off between doing the genome-centric interpretation accurately and doing it conveniently.

Applicability of the annotation approach developed in this study

Overall speaking, in the interpretation of MAGs on the function niche of cellulose hydrolysis, the results returned by the annotation approach developed in this study is of good resolution and reliability. Only these genomes assigned into Group I-a, Group I-b, Group I-c, Group I-d and Group I-e are of potential cellulolytic microorganisms. And among these five groups, genomes of Group I-a and Group I-b correspond to cellulolytic microbes with cellulosome complexes. Genomes of Group I-d are of cellulolytic microbes without cellulosome complexes, and the SLH-cCBM genes might play essential roles in facilitating the CEM complex formation for microbes in this group. Genomes of Groups I-c and Group I-e might be cellulolytic, while the uncertainty comes not from whether they may harbor alternative microbe-cellulose adhesion machineries that could not be recognized by this pipeline. And many well-known cellulose degraders have no well-known mechanisms, e.g., *Cytophaga hutchinsonii*, an aerobic soil cellulolytic organism and *Fibrobacter succinogenes*, an anaerobic ruminal bacterium [15, 49, 51–54]

Conclusion

In summary, this is the first time that a pipeline was developed for a more reliable genome-centric interpretation on the function niche of cellulose hydrolysis. Results of this study suggested the necessity of a finer classification of the cellulosome gene clusters, and not all cellulosome are cell surface anchoring. Enzymes encoded by genes harboring both the cell-surface anchoring SLH module and the

cellulose-binding cCBM module may act as an alternative machinery facilitating the formation of the CEM complexes. The potential cellulose hydrolyzing microbes could be categorized into 5 groups according to the varying synergy mechanisms among the carbohydrate active modules/genes annotated. Pairing with the dbCAN annotation platform, this pipeline is very efficient in identifying potential cellulose hydrolysers by interpreting the complete genomes or MAGs recovered through high-throughput sequencing.

Methods

5243 GenBank Format (GBK) files corresponding to 2786 prokaryote with complete genomes were downloaded from the NCBI genomes_FTP_site (ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_genbank/Bacteria/). The reason why this old archive collection (last updated on Dec. 2nd, 2015) was chosen in this study was that, comparing with the most recently updated archive, this collection had a higher portion of complete genomes from strains whose phenotypes have been well characterized; and the documented phenotypes make it possible to evaluate the reliability of the genome-centric prediction on the corresponding microbes' cellulolytic capability. Another batch of 7904 reference genomes were also downloaded from NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/>) (updated on February, 2019), and there are metagenome assembled genomes (MAGs) among these 7904 reference genomes. These 7904 reference genomes were used to evaluate the applicability of the pipeline in the batch annotation of a large number of MAGs. A detailed summary on these 7904 reference genomes could be found in Appendix file 5.

Fasta Amino Acid sequences (FAA) of the coding regions (often abbreviated as CDS) were extracted from the GBK files with a python script. The FAA files were then subjected to the dbCAN HMMsearch for the CAZy module annotation, following the HMMsearch criteria (e.g. cutoff value) recommended by the dbCAN developers [6]. CAZy (carbohydrate active enzymes) modules were identified in 3898 of these FAA files that corresponded to 2642 prokaryotic strains. The assembly accession numbers and taxonomy affiliation of these 2642 strains have been summarized in the Appendix file 6. As the chromosome and the plasmid in one same microbe strain have separate FAA files, results of the annotation of those separate FAA files of the chromosome and the plasmid in one same microbe strain would be aggregated to represent all the CAZy modules annotated in one microbe strain.

The GH modules that were relevant in the cellulose hydrolysis were classified and read as the exoglucanase GH modules, the endoglucanase GH modules, the xylanase GH modules and the glucosidase GH modules, respectively (Table S2). The CBM modules were classified and read as the cellulose-binding cCBM modules, the hemicellulose-binding CBM modules and other CBM modules, respectively (Table S3). The dockerin, cohesion and the SLH modules were the three important accessory modules in the cellulosome gene clusters. Based on the survey of the carbohydrate active genes in the 2642 complete genomes, frequencies of the CAZy modules co-occurring with one another in same genes were calculated; and the principles applied in such calculation could be found in the supporting information.

Applying the `genoplots` package in R [55], the CAZy module arrangement along genes in genomes could be visualized. Batch visualization of the arrangement of the CAZy modules along all the carbohydrate active genes annotated in each complete genome or MAG could be achieved. Scripts of the pipeline and workflow of the pipeline have been well documented in Github. In addition to the interpretation of the complete genomes from the 2642 CAZy-harboring strains, and the 7904 reference genomes downloaded from NCBI, the pipeline developed in this study was also applied in the annotation of 17 metagenome assembled genomes (MAGs) recovered from a cellulose converting consortia enriched in our previous study [56]. These 17 MAGs can be applied as an example dataset to work with, and all the raw data and results generated on these 17 MAGs have also been deposited in the Github.

List Of Abbreviations

SLH: Surface Layer Homology;

CBM: Carbohydrate Binding Module;

cCBM: cellulose-binding CBM module;

CEM complex: Cellulose-Enzyme-Microbe complex;

MAGs: Metagenome Assembled Genomes (MAGs);

NGS: Next Generation Sequencing;

CAZy: Carbohydrate Active enzyme;

PUL: Polysaccharide-Utilization Loci;

EPS: Extracellular Polymeric Substances.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files and the appendix files. All scripts written in this study are available in <https://github.com/yuboer/genome-centric-portrait-of-cellulose-hydrolysis>.

Competing interests

The authors declare no conflict of interest.

Funding

This work was financially supported by National Key R&D Program of China (grant No. 2018YFC0310600).

Authors' contributions

YW conceived the study, analyzed the data and wrote the manuscript. LL contributed resources of the 2642 complete genomes and the corresponding metadata collection; YX contributed in the CAZy modules annotation. FJ helped with part of the revision. TZ supervised the study. All authors edited the manuscript and approved the final draft.

Acknowledgments

YW wish to thank the University of Hong Kong for the postgraduate scholarship. LL and YX acknowledge the postdoc scholarship provided by the University of Hong Kong.

References

1. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hemsdorf AW, Amano Y, Ise K *et al*: **A new view of the tree of life**. *Nat Microbiol* 2016, **1**:16048.
2. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW: **Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life**. *Nat Microbiol* 2017, **2**(11):1533-1542.
3. Rastogi G, Muppidi GL, Gurram RN, Adhikari A, Bischoff KM, Hughes SR, Apel WA, Bang SS, Dixon DJ, Sani RK: **Isolation and characterization of cellulose-degrading bacteria from the deep subsurface of the Homestake gold mine, Lead, South Dakota, USA**. *J Ind Microbiol Biotechnol* 2009, **36**(4):585-598.
4. Zuzana Mladenovska IMM, Birgitte K. Ahring: **Isolation and characterization of *Caldicellulosiruptor lactoaceticus* sp. nov., an extremely thermophilic, cellulolytic, anaerobic bacterium**. *Arch Microbiol* 1994, **163**:223-230.
5. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics**. *Nucleic Acids Res* 2009, **37**(Database issue):D233-238.
6. 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):W445-451.

7. Barrett K, Hunt CJ, Lange L, Meyer AS: **Conserved unique peptide patterns (CUPP) online platform: peptide-based functional annotation of carbohydrate active enzymes.** *Nucleic Acids Res* 2020, **48**(W1):W110-W115.
8. Uchida K, Jang MS, Ohnishi Y, Horinouchi S, Hayakawa M, Fujita N, Aizawa S: **Characterization of *Actinoplanes missouriensis* spore flagella.** *Appl Environ Microbiol* 2011, **77**(7):2559-2562.
9. Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, Gibson R, Lee HM, Dubois J, Qiu D, Hitti J *et al*: **Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*.** *J Bacteriol* 2001, **183**(16):4823-4838.
10. Bayer EA, Belaich JP, Shoham Y, Lamed R: **The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides.** *Annu Rev Microbiol* 2004, **58**:521-554.
11. Sreekumar S, Baer ZC, Pazhamalai A, Gunbas G, Grippo A, Blanch HW, Clark DS, Toste FD: **Production of an acetone-butanol-ethanol mixture from *Clostridium acetobutylicum* and its conversion to high-value biofuels.** *Nat Protoc* 2015, **10**(3):528-537.
12. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS: **Microbial Cellulose Utilization: Fundamentals and Biotechnology.** *Microbiology and Molecular Biology Reviews* 2002, **66**(3):506-577.
13. Gaston Courtade RW, Asmund K. Rohr, Marita Preims, Alfons K.G. Felice, Maria Dimarogona, Gustav Vaajie-Kolstad, Morten Sorlie, Mats Sandgren, Roland Ludwig, Vincent G.H. Eijsink, and Finn Lillelund Aachmann: **Interactions of a fungal lytic polysaccharide monooxygenase with β -glucan substrates and cellobiose dehydrogenase.** *P Natl Acad Sci USA* 2016, **113**:5922-5927.
14. Sebastian GA Konietzny PBP, Aaron Weimann and Alice C McHardy: **Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders.** *Biotechnol Biofuels* 2014, **7**(124).
15. Wilson DB: **Three microbial strategies for plant cell wall degradation.** *Ann N Y Acad Sci* 2008, **1125**:289-297.
16. Doi RH, Kosugi A: **Cellulosomes: plant-cell-wall-degrading enzyme complexes.** *Nat Rev Microbiol* 2004, **2**(7):541-551.
17. Akin DE, Lyon, C.E. , Windham, W.R. , and Rigsby, L.L. : **Physical degradation of lignified stem tissues by ruminal fungi.** *Applied and Environmental Microbiology* 1989, **55**:611-616.
18. Akin DE, Borneman, W.S., and Lyon, C.E.: **Degradation of leaf blades and stems by monocentric and polycentric isolates of ruminal fungi.** *Animal Feed Science and Technology* 1990, **31**:205-221.
19. Tengerdy RP, W. H. Rho, and A. M. Mohagheghi: **Liquid fluidized bed starter culture of *Trichoderma reesei* for cellulase production.** *Applied Biochemistry and Biotechnology* 1991, **27**:195–204.
20. Busto MD, Ortega, N., and Perez-Mateos, M. : **Location, kinetics and stability of cellulases induced in *trichoderma reesei* cultures.** *Bioresour Technol* 1996, **57**:187-192.
21. Ho YW, and Abdullah, N.: **The role of rumen fungi in fibre digestion: Review.** *Asian-Australas J Anim Sci* 1999, **12**:104-112.

22. Kudo H, K.-J. Cheng, and J. W. Costerton: **Electron microscopic study of the methylcellulose-mediated detachment of cellulolytic rumen bacteria from cellulose fibers.** *Can J Microbiol* 1987, **33**:267–271.
23. Bayer EA, R. Kenig, and R. Lamed: **Adherence of Clostridium thermocellum to cellulose.** *Journal of Bacteriology* 1983(156):818-827.
24. Raphael Lamed ES, Edward A. Bayer: **Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum.** *Journal of Bacteriology* 1983, **156**(2):828-836.
25. Akihiko Kosugi KM, Yutaka Tamaru, and Roy H. Doi: **Cell-Surface-Anchoring Role of N-Terminal Surface Layer Homology Domains of Clostridium cellulovorans EngE.** *J Bacteriol* 2002, **184**(4):884-888.
26. Doi YTaRH: **Three Surface Layer Homology Domains at the N Terminus of the Clostridium cellulovorans Major Cellulosomal Subunit EngE.** *J Bacteriol* 1999, **181**(10):3270-3276.
27. Miron J, and C. W. Forsberg: **Characterisation of cellulose-binding proteins that are involved in the adhesion mechanism of Fibrobacter intestinalis DR7.** *Appl Microbiol Biotechnol* 1999, **51**:491–497.
28. Terrapon N, Lombard V, Drula E, Lapebie P, Al-Masaudi S, Gilbert HJ, Henrissat B: **PULDB: the expanded database of Polysaccharide Utilization Loci.** *Nucleic Acids Res* 2018, **46**(D1):D677-D683.
29. Feinberg L, Foden J, Barrett T, Davenport KW, Bruce D, Detter C, Tapia R, Han C, Lapidus A, Lucas S *et al*: **Complete genome sequence of the cellulolytic thermophile Clostridium thermocellum DSM1313.** *J Bacteriol* 2011, **193**(11):2906-2907.
30. Ellis LD, Holwerda EK, Hogsett D, Rogers S, Shao X, Tschaplinski T, Thorne P, Lynd LR: **Closing the carbon balance for fermentation by Clostridium thermocellum (ATCC 27405).** *Bioresour Technol* 2012, **103**(1):293-299.
31. Izquierdo JA, Goodwin L, Davenport KW, Teshima H, Bruce D, Detter C, Tapia R, Han S, Land M, Hauser L *et al*: **Complete Genome Sequence of Clostridium clariflavum DSM 19732.** *Stand Genomic Sci* 2012, **6**(1):104-115.
32. Giallo J. GC, and Belaich JP: **Metabolism and Solubilization of Cellulose by Clostridium cellulolyticum H10.** *Applied and Environmental Microbiology* 1985, **49**:1216-1221.
33. Li LL, Taghavi S, Izquierdo JA, van der Lelie D: **Complete genome sequence of Clostridium sp. strain BNL1100, a cellulolytic mesophile isolated from corn stover.** *J Bacteriol* 2012, **194**(24):6982-6983.
34. Lee D. Mermelstein NEW, George N. Bennett, and Eleftherios T. Papoutsakis: **Expression of cloned homologous fermentative genes in Clostridium acetobutylicum ATCC 824.** *Nature Biotechnology* 1992, **10**:190-195.
35. F.A. Rainey AMD, P.H. Janssen, D. Saul, A. Rodrigo, P.L. Bergquist, R.M. Daniel, E. Stackebrandt and H.W. Morgan: **Description of Caldicellulosiruptorsaccharolyticus gen. nov., sp. nov: An obligately anaerobic, extremely thermophilic, cellulolytic bacterium.** *FEMS Microbiology Letters* 1994, **120**:263-266.
36. Gibbs MD, Reeves RA, Farrington GK, Anderson P, Williams DP, Bergquist PL: **Multidomain and multifunctional glycosyl hydrolases from the extreme thermophile Caldicellulosiruptor isolate Tok7B.1.** *Curr Microbiol* 2000, **40**(5):333-340.

37. Miroshnichenko ML, Kublanov IV, Kostrikina NA, Tourova TP, Kolganova TV, Birkeland NK, Bonch-Osmolovskaya EA: **Caldicellulosiruptor kronotskyensis sp. nov. and Caldicellulosiruptor hydrothermalis sp. nov., two extremely thermophilic, cellulolytic, anaerobic bacteria from Kamchatka thermal springs.** *Int J Syst Evol Microbiol* 2008, **58**(Pt 6):1492-1496.
38. Hamilton-Brehm SD, Mosher JJ, Vishnivetskaya T, Podar M, Carroll S, Allman S, Phelps TJ, Keller M, Elkins JG: **Caldicellulosiruptor obsidiansis sp. nov., an anaerobic, extremely thermophilic, cellulolytic bacterium isolated from Obsidian Pool, Yellowstone National Park.** *Appl Environ Microbiol* 2010, **76**(4):1014-1020.
39. Lochner A, Giannone RJ, Keller M, Antranikian G, Graham DE, Hettich RL: **Label-free quantitative proteomics for the extremely thermophilic bacterium Caldicellulosiruptor obsidiansis reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass.** *J Proteome Res* 2011, **10**(12):5302-5314.
40. Kanafusa-Shinkai S, Wakayama J, Tsukamoto K, Hayashi N, Miyazaki Y, Ohmori H, Tajima K, Yokoyama H: **Degradation of microcrystalline cellulose and non-pretreated plant biomass by a cell-free extracellular cellulase/hemicellulase system from the extreme thermophilic bacterium Caldicellulosiruptor bescii.** *J Biosci Bioeng* 2013, **115**(1):64-70.
41. Basen M, Rhaesa AM, Kataeva I, Prybol CJ, Scott IM, Poole FL, Adams MW: **Degradation of high loads of crystalline cellulose and of unpretreated plant biomass by the thermophilic bacterium Caldicellulosiruptor bescii.** *Bioresour Technol* 2014, **152**:384-392.
42. Pason P, Kyu KL, Ratanakhanokchai K: **Paenibacillus curdlanolyticus strain B-6 xylanolytic-cellulolytic enzyme system that degrades insoluble polysaccharides.** *Appl Environ Microbiol* 2006, **72**(4):2483-2490.
43. Wang CM, Shyu CL, Ho SP, Chiou SH: **Characterization of a novel thermophilic, cellulose-degrading bacterium Paenibacillus sp. strain B39.** *Lett Appl Microbiol* 2008, **47**(1):46-53.
44. Lal S, Tabacchioni S: **Ecology and biotechnological potential of Paenibacillus polymyxa: a minireview.** *Indian J Microbiol* 2009, **49**(1):2-10.
45. Waeonukul R, Kyu KL, Sakka K, Ratanakhanokchai K: **Isolation and characterization of a multienzyme complex (cellulosome) of the Paenibacillus curdlanolyticus B-6 grown on Avicel under aerobic conditions.** *J Biosci Bioeng* 2009, **107**(6):610-614.
46. Asha BM: **Purification and Characterization of a Thermophilic Cellulase from a Novel Cellulolytic Strain, Paenibacillus barcinonensis.** *Journal of Microbiology and Biotechnology* 2012, **22**(11):1501-1509.
47. Park IH, Chang J, Lee YS, Fang SJ, Choi YL: **Gene cloning of endoglucanase Cel5A from cellulose-degrading Paenibacillus xylanilyticus KJ-03 and purification and characterization of the recombinant enzyme.** *Protein J* 2012, **31**(3):238-245.
48. Poehlein A, Zverlov VV, Daniel R, Schwarz WH, Liebl W: **Complete Genome Sequence of Clostridium stercorarium subsp. stercorarium Strain DSM 8532, a Thermophilic Degradar of Plant Cell Wall Fibers.** *Genome Announc* 2013, **1**(2):e0007313.

49. Wilson DB: **Microbial diversity of cellulose hydrolysis.** *Curr Opin Microbiol* 2011, **14**(3):259-263.

50. Weimer PJ, Price NP, Kroukamp O, Joubert LM, Wolfaardt GM, Van Zyl WH: **Studies of the extracellular glycocalyx of the anaerobic cellulolytic bacterium Ruminococcus albus 7.** *Appl Environ Microbiol* 2006, **72**(12):7559-7566.

51. Ransom-Jones E, Jones DL, McCarthy AJ, McDonald JE: **The Fibrobacteres: an important phylum of cellulose-degrading bacteria.** *Microb Ecol* 2012, **63**(2):267-281.

52. Zhu Y, McBride MJ: **The unusual cellulose utilization system of the aerobic soil bacterium Cytophaga hutchinsonii.** *Appl Microbiol Biotechnol* 2017, **101**(19):7113-7127.

53. Burnet MC, Dohnalkova AC, Neumann AP, Lipton MS, Smith RD, Suen G, Callister SJ: **Evaluating Models of Cellulose Degradation by Fibrobacter succinogenes S85.** *PLoS One* 2015, **10**(12):e0143809.

54. Wilson DB: **Cellulases and biofuels.** *Curr Opin Biotechnol* 2009, **20**(3):295-299.

55. Guy L, Kultima JR, Andersson SG: **genoPlotR: comparative gene and genome visualization in R.** *Bioinformatics* 2010, **26**(18):2334-2335.

56. Xia Y, Wang Y, Wang Y, Chin FY, Zhang T: **Cellular adhesiveness and cellulolytic capacity in Anaerolineae revealed by omics-based genome interpretation.** *Biotechnol Biofuels* 2016, **9**:111.

Figures

Figure 1

Co-occurrence frequencies among the CAZy modules Note: Vertically listed were the twenty-one selected CAZy modules, including the three exoglucanase GH modules, the eight endoglucanase GH modules, the seven xylanase GH modules, the cohesin, the dockerin and the SLH module. The CAZy modules lining horizontally were those modules being observed in same genes with at least one of the vertically listed

CAZy modules. The scale bar on the right presented the co-occurring frequencies (x) in the log format of $\lg(x+0.01)$, and the plain number 'x' was summarized in the Appendix file 2.

Figure 2
Illustration on the clustering patterns of the CAZy modules along gene

Figure 3

Circle tree of the 2642 genomes assigned into different subgroups with varying cellulolytic competency. The taxonomy levels of Kingdom (Bacteria), phylum, class, order, family, genus and strain were presented by the successive inner nodes. Genomes assigned to the first 5 subgroups of Group I are highlighted in 5 colors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Appendixfile6.xlsx](#)
- [Appendixfile5.tab](#)
- [Appendixfile4.csv](#)
- [Appendixfile3.xlsx](#)
- [Appendixfile2.xlsx](#)
- [Appendixfile1.xlsx](#)

- [SupportingInformation.docx](#)