Sift-PULs: A public repository for specific functional polysaccharide utilization loci

Tao Song[1]*, Congchong Wei[1], Dezhi Yuan[2], Shengwei Xiang[1], Lin Liu[1], Hua Lv[1], Binkai Wang[1], Ting Huang[1], Kelei Zhao[1], Xinrong Wang[1], Yiwen Chu[1], Jiafu Lin[1]*

1: Antibiotics Research and Re-evaluation Key Laboratory of Sichuan Province, Sichuan Industrial Institute of Antibiotics, School of pharmacy, Chengdu University, 610106, Chengdu, China

2: Moutai Institute, Renhuai 564500, Guizhou Province, China

*: Corresponding authors of this paper

Corresponding author:

Tao Song and Jiafu Lin

Sichuan Industrial Institute of Antibiotics,

Chengdu University, Chengdu, 610106, P. R. China

Tel: +86-028-84216083, Fax: +86-028-84216083

E-mail: songtao@cdu.edu.cn, linjiafu@cdu.edu.cn

Abstract

Background

Polysaccharide utilization loci (PULs) were bacterial gene clusters encoding genes responsible for polysaccharide utilization process. PUL studies are blooming in recent years but the biochemical characterization speed is relative slow. There is a growing demand for PUL database with function annotations.

Results

Using signature genes corresponding for specific polysaccharide, 10422 PULs specific for 6 polysaccharides (agar, alginate, pectin, carrageenan, chitin and β-manan) from various bacterial phyla were predicted. Then online website of specific functional polysaccharide utilization loci (Sift-PULs) was constructed. Sift-PULs provides a repository where users could browse, search and download interested PULs without registration.

Conclusions

The key advantage of Sift-PULs is to assign a function annotation of each PUL, which is not available in existing PUL databases. PUL's functional annotation lays a foundation for studying novel enzymes, new pathways, PUL evolution or bioengineering. The website is available on http://sift-puls.org

Keywords: Polysaccharide utilization loci; Database; Function annotation

Introduction

PUL s(polysaccharide utilization loci) are bacterial gene clusters that encoding a variety of functional genes responsible for the transcription, degradation, transport and metabolism of polysaccharides (Grondin *et al.*, 2017). Since the discovery of starch PUL (also named as Sus. starch utilization system) from *Bacteroides thetaiotaomicron*, more and more PULs have been found from different ecosystems and various bacteria phyla (D'Elia and Salyers, 1996; Foley *et al.*, 2016; Chen *et al.*, 2018; Despres *et al.*, 2016; Grondin *et al.*, 2017). Discovered PULs are found to target at many different types of polysaccharides, including xylan, β-mannan or pectin (Tang *et al.*, 2017; Bagenholm *et al.*, 2017; Reddy *et al.*, 2016; Ficko-Blean *et al.*, 2017; Despres *et al.*, 2016; Pluvinage *et al.*, 2018). Now PUL studies are becoming a hotspot because their significant importance in ecology, evolution and bioengineering(Grondin *et al.*, 2017). Although PULs have important biological functions, biochemical identification in laboratory is too slow, resulting in number scarcity and hindering the progress of PUL study.

In view of the sparse data of PULs and PULs' significant biological functions, it is very necessary to use bioinformatics way to identify PULs. There are currently three PUL databases available, including PULDB, CGCs and dbCAN-PUL(Terrapon *et al.*, 2015; Zhang *et al.*, 2018; Ausland *et al.*, 2021). PULDB is the first PUL database, which uses SusC/D gene pair and carbohydrate active enzymes for prediction. It mainly includes PULs from *Bacteroidetes*. CGCs predict PULs using transcription factors, transport

proteins and carbohydrate active enzymes while dbCAN-PUL does not predict PULs but provides experimentally confirmed PULs. It is worth mentioning that although these three databases provide PUL collections, there is no annotation for predicted PULs. So researchers who want to find PULs targeting at specific polysaccharide need to manually check the predicted PULs. This is very laborious. PUL with a function annotation is helpful for researchers to answer new hypotheses and provide a basis for new discoveries. With the increase of PUL studies, the requirement of PUL database with specific function annotations has become more and more urgent. Unfortunately, there is no PUL database providing specific function annotation now.

PULDB and CGCs are the two main databases currently used for PUL prediction, and no functional prediction is given for the predicted PULs. Therefore, we used signature genes corresponding to 6 different polysaccharides (agar, alginate, pectin, carrageenan, chitin and β-manan) to predict PULs, and gave function predictions. Then Sift-PULs website is constructed, where users could easily search and download interested Sift-PULs. Sift-PULs serves as repository for researchers who focus on one specific polysaccharide and need large scale data to discover novel protein, utilization pathways or evolutional process.

Materials and methods

Data retrieval

Bacterial genomes were mainly downloaded from NCBI database (download was finished in 2021.03.01). Genomes at different assemble levels (contig, scaffold, complete or chromosome) were downloaded using a home-made script. Only GBFF format file were retrieved from FTP link.

Data normalization

The genbank file of a bacterial genome was parsed using Biopython package, then protein sequences within bacterial genome were extracted into a single fasta file. The name of each protein is normalized into following format: GCF number of genome, contig name, serial number on the contig, gene start position, gene end position and gene direction. Therefore, protein name was a unique signature which contained essential information for prediction.

Selection of signature gene

In total, PULs that were specific for alginate, agar, carrageenan, chitin, β-mannan and pectin (polygalacturonicacid) were considered in this manuscript. In this study, signature genes were classified into two categories, core genes and alternative genes (Supplementary material 1). Core genes referred to genes that were essential for the polysaccharide utilization process, including ones responsible for monosaccharide metabolism (e.g. unique 3,6 anhydro-L galactose metabolic genes for agar) or unique metabolic process (e.g. GH130 mannobiose phosphorylase for mannan utilization). Core genes were determined if they were commonly found in most biochemically PULs. Alternative genes were usually carbohydrate active enzymes responsible for polysaccharide utilization. Alternative genes were

92 determined if they appeared in characterized PULs or their activities were related to polysaccharide
93 degradation.

94 Hmmer model build

95 Most hmmer models responsible for signature genes were built locally. To build an hmmer model,
96 experimentally validated protein sequences were first collected and aligned using MUSCLE(Edgar, 2004),
97 followed by manual correction. Proteins with experimental evidences from CAZYs and Unipro were used
98 as test data to test the true positive rate and false positive rate. The re-build hmmer model should
99 have >95% true positive rate and <5% false positive rate under a specific threshold. When the signature
100 gene was from a family with only one enzyme activity and this family had very few experimentally
101 confirmed members (less than 3), the corresponding hmmer model was retrieved from Pfam and dbCAN
102 (Finn *et al.*, 2013; Zhang *et al.*, 2018).

103 Sift-PULs prediction

104 Sift-PUL prediction needed 6 steps:

105 (1) Firstly, normalized protein fasta file was analyzed using Hmmer against corresponding models.
106 If domain of a gene was the same as core gene or alternative gene, it was recorded as core
107 gene or alternative gene, respectively.
108 (2) Then, we investigated whether it was possible to put core genes and at least one alternative
109 gene into a gene cluster with less than 50 genes. If did, serial numbers of matched core genes or
110 alternative genes were recorded.
111 (3) Minimum PUL was defined as gene cluster contained minimum members including all core
112 genes and at least one alternative gene. Extended the minimum PUL to both sides until the gene
113 number reached 50. Then extended PUL was defined as maximum PUL.
114 (4) Calculate the frequency of individual domain in minimum PULs. Domains with >10% frequency
115 were defined as high frequency domain.
116 (5) PUL boundary of minimum PUL was extended until the adjacent and consecutive 5 genes did
117 not have high frequency domain. If the extended PULs were smaller than maximum PUL,
118 extended boundary was used. Otherwise, maximum PUL boundary was used.

119 Online database construct

120 The website of Sift-PULs was constructed using Vue.js (javascript) and Django (python). Database
121 is implemented using PosgreSQL.

122 **Result and discussion**

123 **Data collections of Sift-PULs**

124 PULs are bacterial gene clusters that have essential biological functions. Considering increasing
125 interest of researchers in PUL study and PULs' slow identification in laboratory, it is necessary to

establish a PUL database with function prediction. Using signature genes that specific to corresponding polysaccharide, 10422 PULs were identified, including 2347 pectin PULs, 1140 manan PULs, 1938 alginate PULs, 4723 chitin PULs, 186 agar PULs and 88 carrageenan PULs (Fig 1A). Meanwhile, predicted PULs came from different phyla including *Proteobacteria* (4140 PULs), *Firmicutes* (3335 PULs), *Bacteroidetes* (2342 PULs) and *Actinobacteria* (537 PULs). Noteworthy, Sift-PULs showed a potent as a reference database for discovering novel PULs. For example, predicted carrageenan PULs came from 5 phyla (*Actinobacteria*, *Bacteroidetes*, *Planctomycetes*, *Proteobacteria* and *Firmicutes*), and now only carrageenan from *Bacteroidetes* was experimentally verified. Moreover, bacterial genomes above contig level were used for sift-PULs prediction in this study, therefore most predicted sift-PULs come from bacteria had contig or scaffold genomes (Fig 1B).

In current study, hmmer models were locally built to ensure signature genes' specificity, then combination of signature genes was used for PUL's function prediction. Still, it was possible that our predicted results may contain false positives.

To investigate the data reliability of predicted sift-PULs, firstly, we tried to evaluate the prediction method to give hint about data accuracy. However, it did not succeed because of insufficient number of experimentally confirmed PULs. For example, there was only one report of carrageenan PUL, 2 reports of agar PULs, less than 10 reports of pectin PULs. Then, we tried to find evidence in database containing bacterial polysaccharide utilization information (biodive). However, bacteria with sift-PULs were either not recorded in biodive database, or corresponding polysaccharide utilization information was not recorded. Matched results were too few to get any useful conclusions. In the end, we focused on bacteria with agar-PULs. This was because agar was commonly used in bacterial cultivation for almost 100 years, agar degradation phenotype could be easily seen on plate, and this information was more likely to be recorded in literature. Surprisingly, 70 out of 186 bacteria with predicted agar PULs could degrade agar and the rest were not mentioned (Supplementary material 2). This implied the predicted agar PULs were relative reliable. The accuracy of agar PULs also indicated sift-PULs could be used as reference databank for researchers.

Comparison with existing PUL databases

Using signature genes to predict PUL was commonly in current research(Terrapon *et al.*, 2015; Zhang *et al.*, 2018). For example, SIFT-PULS used the PUL conservative SusC/D gene pair and carbohydrate active enzymes, and CGCs used transcription factors, sugar transporters and carbohydrate active enzymes. The signature genes used in prediction determine the properties of the obtained PUL. For example, the PULs in SIFT-PULS only came from *Bacteroidetes*, because the SusC/D gene pair was mainly derived from *Bacteroidetes*. Because the selected signature genes are not specific to polysaccharides in PULDB or CGCs, none of these two databases could give function prediction. The signature genes used in Sift-PULs in this article were specific to each polysaccharide, therefore function annotation was possible. The function prediction greatly reduced the workload of researchers searching for the corresponding function PUL.

Sift-PULs included 10422 PULs, which less than with 43156 PULs in PULDB (Table 1). This was probably because sift-PULs only focus on 6 polysaccharides. Meanwhile, PULs from Sift -PULS were

from multiple bacterial phyla. This was similar to CGCs but different with sift-PULS, in which only *Bacteroidetes* was considered. Compared with sift-PULs and CGCs, the most important feature of Sift-PULs was that it could give function predictions.

Web interface

At the start page of sift-PULs, there were six sections: home, search, browse, download, links and help (Fig 1A). At home page, there was a brief introduction of sift-PULs, where users could quickly learn how sift-PULs were predicted. Important update would also be showed here. User could find interested sift-PULs in two ways. First, in the search section, users could search for interested PULs using different keywords, for example polysaccharide name, taxid, GCF number, phylum name, species name or protein domain name (Fig 1B). Second, in browse section, sift-PULs were classified by polysaccharide or phyla (Fig 1C). By clicking the search button in search section or links in browse section, interested sift-PULs would be displayed (Fig 1D). After clicking 'view' button interested PUL, PUL information would be displayed in a pop-up page, which contains the PUL information, download option, gene cluster map and gene information(Fig 1E)..

Sift-PULs also provide batch download service, which was convenient for users who required large amount of data. In download section, users could easily download all sift-PULs data (Fig 1F). There were three format files available, including a genbank file that included complete DNA sequence of PUL, DNA fasta file that included DNA sequences for individual CDS, protein fasta file that included protein sequences for individual CDS. Users could download these files when browsing individual PUL.

Conclusions

Sift-PULs website provides a public repository where users could easily access, search and download PULs with specific function annotation, which helps researchers to build a local database and come up with novel hypothesis. For example, Sift-PULs could help biochemists discover novel enzymes (study proteins that are not characterized but have high frequency score) and find novel degradation pathways. In future, Sift-PULs would update once a year. Update would include sift-PULs from newly sequenced genomes or sift-PULs targeting at new polysaccharides (e.g. α-mannan, starch and ulvan). Online prediction service of sift-PULs is also under construction.

Declaration

-Abbreviations (if applicable)

PULs: polysaccharide utilization loci

-Ethics approval and consent to participate

Not applicable

-Consent for publication

Not applicable

-Availability of data and materials

Not applicable

-Funding

This work was supported by the Sichuan Science and Technology Program (Grant Number: 2019YJ0282); Collaboration and Innovation on New Anti-biotic Development and Industrialization（Grant Number: 2016-XT00-00023-G); Major New Drugs Innovation and Development (Grant Number: 2018ZX09711-001); The talent Introduction of Project of Chengdu University (2081918021) and Key Laboratory of Coarse Cereal Processing, Ministry of Agriculture and Rural Affairs (2020CC009).

Publishing cost was funded by Antibiotics Research and Re-evaluation Key Laboratory of Sichuan Province (Grant Number: ARRLKF19-01, ARRLKF19-05); The Science and technology projects from Department of Ecology and Environment of Sichuan Province (Grant No. 2019HB16).

-Competing Interests

We declare that we have no competing interest.

-Authors' contributions

TS: conceptualization, writing- reviewing and editing; CW and DY: website construction and maintain; SX, LL and H: visualization, development or design of methodology; BW and XZ: data analysis and visualization; XZ and TH: data curation; KZ: project administration; XW: funding acquisition; YC: funding acquisition and writing- reviewing; JL: conceptualization, writing- reviewing and editing, supervision. All authors have read and approved the manuscript.

-Acknowledgements

We thank Dr. Jan Hendrik Hehemann (Max Planck Institute for Marine Microbiology) for his critical comments on polysaccharide utilization pathways.

Reference

Ausland,C. *et al.* (2021) dbCAN-PUL: a database of experimentally characterized CAZyme gene clusters and their substrates. *Nucleic Acids Res.*, **49**, D523–D528.

Bagenholm,V. *et al.* (2017) Galactomannan catabolism conferred by a polysaccharide utilisation locus of Bacteroides ovatus : enzyme synergy and crystal structure of a β-mannanase. *J. Biol. Chem.*, **292**, 229–243.

Chen,J. *et al.* (2018) Alpha‐ and beta‐mannan utilization by marine Bacteroidetes. *Environ. Microbiol.*, **20**, 4127–4140.

D'Elia,J.N. and Salyers,A.A. (1996) Effect of regulatory protein levels on utilization of starch by Bacteroides thetaiotaomicron. *J. Bacteriol.*, **178**, 7180–7186.

Despres,J. *et al.* (2016) Xylan degradation by the human gut Bacteroides xylanisolvens XB1AT involves two distinct gene clusters that are linked at the transcriptional level. *BMC Genomics*, **17**, 326.

233 Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput.
234 *Nucleic Acids Res.*, **32**, 1792–1797.

235 Ficko-Blean,E. *et al.* (2017) Carrageenan catabolism is encoded by a complex regulon in marine
236 heterotrophic bacteria. *Nat. Commun.*, **8**, 1685.

237 Finn,R.D. *et al.* (2013) Pfam: The protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

238 Foley,M.H. *et al.* (2016) The Sus operon: a model system for starch uptake by the human gut
239 Bacteroidetes. *Cell. Mol. Life Sci.*, **73**, 2603–2617.

240 Grondin,J.M. *et al.* (2017) Polysaccharide utilization loci: Fueling microbial communities. *J. Bacteriol.*,
241 **199**, e00860-16.

242 Pluvinage,B. *et al.* (2018) Molecular basis of an agarose metabolic pathway acquired by a human
243 intestinal symbiont. *Nat. Commun.*, **9**, 1043.

244 Reddy,S.K. *et al.* (2016) A β‐mannan utilization locus in Bacteroides ovatus involves a GH36
245 α‐galactosidase active on galactomannans. *FEBS Lett.*, **590**, 2106–2118.

246 Tang,K. *et al.* (2017) Characterization of potential polysaccharide utilization systems in the marine
247 Bacteroidetes Gramella flava JLT2011 using a multi-omics approach. *Front. Microbiol.*, **8**.

248 Terrapon,N. *et al.* (2015) Automatic prediction of polysaccharide utilization loci in Bacteroidetes species.
249 *Bioinformatics*, **31**, 647–655.

250 Zhang,H. *et al.* (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.
251 *Nucleic Acids Res.*, **46**, W95–W101.

252

253 Table 1 Comparison Sift-PULDB with PULDB and CGCs，N.A.: not available

	PULDB	CGCs	Sift-PULs
Signature Genes	SusC/D and CAZys	Transcription factor, transporter and CAZys	Specific genes for each polysaccharide
PUL numbers	43156	N.A.	10422
Multiple phyla	No	Yes	Yes
Bacteria assemble level	Complete	Complete	above contig
Function prediction	No	No	yes

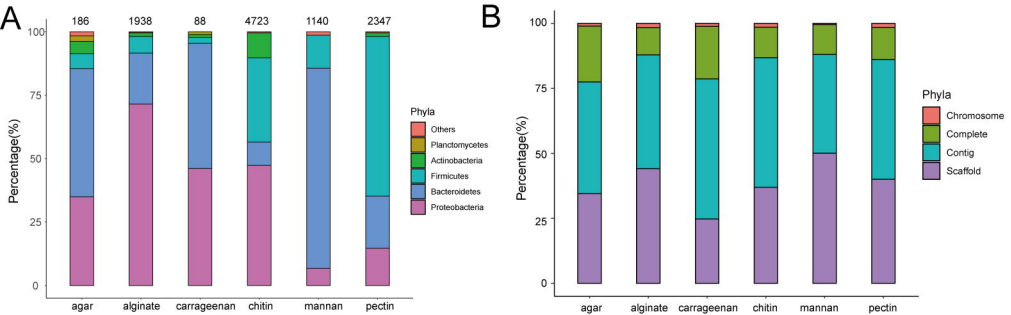
254

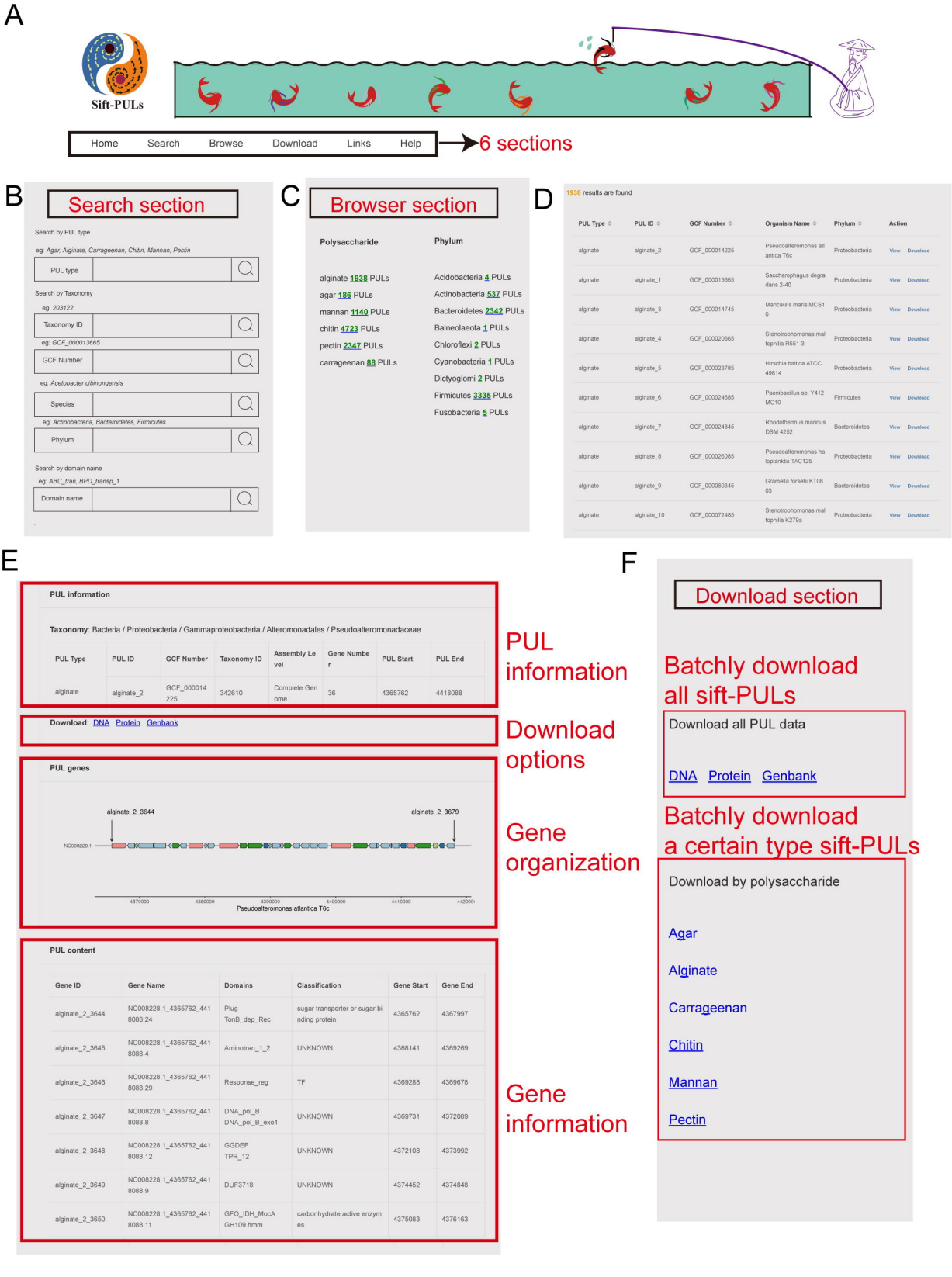
255

Figure legends：

Figure 1 summary of predicted sift-PULs. A: phyla distribution of bacteria with predicted sift-PULs. B: Assemble level of bacteria with predicted sift-PULs.

Figure 2: Screenshots of sift-PULs website. A: Menu in sift-PULs website link to different sections. B: Search section in sift-PULs website. Users could search sift-PULs by phyla, species name, txa id and domain name. C: Browse section in sift-PULs website. Sift-PULs were classified by polysaccharide or phyla. D: Screenshot of sift-PULs list after user click search or browse button. E: Web interface for a sift-PUL using a alginate PUL as an example. F: Web interface for batch download.


A
186
1938
88
4723
1140
2347
100
75
50
25
0
Percentage(%)
agar
alginate
carrageenan
chitin
mannan
pectin
Phyla
Others
Planctomycetes
Actinobacteria
Firmicutes
Bacteroidetes
Proteobacteria
B
100
75
50
25
0
Percentage(%)
agar
alginate
carrageenan
chitin
mannan
pectin
Phyla
Chromosome
Complete
Contig
Scaffold


A
Sift-PULs
Home
Search
Browse
Download
Links
Help
6 sections
B
Search section
Search by PUL type
eg: Agar, Alginate, Carrageenan, Chitin, Mannan, Pectin
PUL type
Search by Taxonomy
eg: 203122
Taxonomy ID
eg: GCF_000013665
GCF Number
eg: Acetobacter cibinongensis
Species
eg: Actinobacteria, Bacteroidetes, Firmicutes
Phylum
Search by domain name
eg: ABC_tran, BPD_transp_1
Domain name
C
Browser section
Polysaccharide
alginate 1938 PULs
agar 186 PULs
mannan 1140 PULs
chitin 4723 PULs
pectin 2347 PULs
carrageenan 88 PULs
Phylum
Acidobacteria 4 PULs
Actinobacteria 537 PULs
Bacteroidetes 2342 PULs
Balneolaeota 1 PULs
Chloroflexi 2 PULs
Cyanobacteria 1 PULs
Dictyoglomi 2 PULs
Firmicutes 3335 PULs
Fusobacteria 5 PULs
D
1938 results are found
PUL Type
PUL ID
GCF Number
Organism Name
Phylum
Action
alginate alginate_2 GCF_000014225 Pseudoalteromonas atlantica T6c Proteobacteria View Download
alginate alginate_1 GCF_000013665 Saccharophagus degradans 2-40 Proteobacteria View Download
alginate alginate_3 GCF_000014745 Maricaulis maris MCS10 Proteobacteria View Download
alginate alginate_4 GCF_000020665 Stenotrophomonas maltophilia R551-3 Proteobacteria View Download
alginate alginate_5 GCF_000023785 Hirschia baltica ATCC 49814 Proteobacteria View Download
alginate alginate_6 GCF_000024685 Paenibacillus sp. Y412MC10 Firmicutes View Download
alginate alginate_7 GCF_000024845 Rhodothermus marinus DSM 4252 Bacteroidetes View Download
alginate alginate_8 GCF_000026085 Pseudoalteromonas haloplanktis TAC125 Proteobacteria View Download
alginate alginate_9 GCF_000060345 Gramella forsetii KT0803 Bacteroidetes View Download
alginate alginate_10 GCF_000072485 Stenotrophomonas maltophilia K279a Proteobacteria View Download
E
PUL information
Taxonomy: Bacteria / Proteobacteria / Gammaproteobacteria / Alteromonadales / Pseudoalteromonadaceae
PUL Type
PUL ID
GCF Number
Taxonomy ID
Assembly Level
Gene Number
PUL Start
PUL End
alginate alginate_2 GCF_000014225 342610 Complete Genome 36 4365762 4418088
Download: DNA Protein Genbank
PUL genes
alginate_2_3644
alginate_2_3679
NC008228.1
4370000
4380000
4390000
4400000
4410000
4420000
Pseudoalteromonas atlantica T6c
PUL content
Gene ID
Gene Name
Domains
Classification
Gene Start
Gene End
alginate_2_3644 NC008228.1_4365762_4418088.24 Plug TonB_dep_Rec sugar transporter or sugar binding protein 4365762 4367997
alginate_2_3645 NC008228.1_4365762_4418088.4 Aminotran_1_2 UNKNOWN 4368141 4369269
alginate_2_3646 NC008228.1_4365762_4418088.29 Response_reg TF 4369288 4369678
alginate_2_3647 NC008228.1_4365762_4418088.8 DNA_pol_B DNA_pol_B_exo1 UNKNOWN 4369731 4372089
alginate_2_3648 NC008228.1_4365762_4418088.12 GGDEF TPR_12 UNKNOWN 4372108 4373992
alginate_2_3649 NC008228.1_4365762_4418088.9 DUF3718 UNKNOWN 4374452 4374848
alginate_2_3650 NC008228.1_4365762_4418088.11 GFO_IDH_MocA GH109.hmm carbonhydrate active enzymes 4375083 4376163
PUL information
Download options
Gene organization
Gene information
F
Download section
Batchly download all sift-PULs
Download all PUL data
DNA Protein Genbank
Batchly download a certain type sift-PULs
Download by polysaccharide
Agar
Alginate
Carrageenan
Chitin
Mannan
Pectin