Supplemental 1: HMM search results against all HMMs for glycoside hydrolase families. The catalytic residues are in red.

```
Query:        Sequence  [L=1545]
Scores for complete sequence (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model        Description
    ------- ------ -----    ------- ------ -----    ---- --  --------     -----------
    2.6e-05   19.0   0.2    4.6e-05   18.2   0.2    1.3  1  GH39
    0.00061   13.8   0.0     0.0012   12.9   0.0    1.4  1  GH79


Domain annotation for each model (and alignments):

top    >> GH39

   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------     ----
   1 !   18.2   0.2   1.9e-07   4.6e-05     149     338 ..     185     320 ..     166     329 .. 0.79

  Alignments for each domain:
  == domain 1    score: 18.2 bits;  conditional E-value: 1.9e-07
  GH39_allsf 149 nggeeVrawvfEiWNEPNltFagfWnsdlswgrgappgdpaqYfeLlkatyraIKavdP......nakVgGPalapapgsdldwlraFLlfckgngvpp 241
                 n+ +++++ E+ NEPNl +++        g+ + d  Y ++++ + a+K v+      +kVgGP l++ ++ + ++F ++++++ +
    Sequence 185 NP--KIEYI--EVGNEPNLE-PAYY--------GHVKDDIPGYMRMYQGMSEAVKWVNAqglagqPLKVGGPVLSG---YNFEKQKQFVDLAYAASYQV 267
                 55..89999..********6.5566........46677777**************998877777779******888...****************** PP

  GH39_allsf 242 DfvswHpYgsgnppedefgyttlpdrpgpadgwlrphprfyfrredvaalrallaayGlpprpPiwiTEyGwstspgtpgyvaarlpdepgvdepgq 338
                 DfvswH Y +        + ++++             ++++ l+++ p +   ++EyGw              +++g + p++
    Sequence 268 DFVSWHRYQED--------------VRMNETQ--------------EIEMKSYLHKF-YPNAI-TIVSEYGW--K------------GGGGLSDPTN 320
                 ********988..............5555552..............35666666666.54455.77888888..6............6666666665 PP


top    >> GH79

   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------     ----
   1 !   12.9   0.0   4.9e-06    0.0012     141     280 ..     184     276 ..     170     279 .. 0.72

  Alignments for each domain:
  == domain 1    score: 12.9 bits;  conditional E-value: 4.9e-06
  GH79_allsf 141 aaGdnLdafElGNEPDLYatsrHghrprgpanw.tpsdYvaewgslataiseaaqleaaaadgnilskgkalligPsGffagsahttnsstvgggWppv 238
                 +  +++ ++E+GNEP+L           pa + ++d+++ + + + +s  ++ + ++++++ +++    ++gP   +                g
    Sequence 184 K-NPKIEYIEVGNEPNLE-----------PAYYgHVKDDIPGYMRMYQGMS---EAVKWVNAQGLAGQPL-KVGGPV---LS-----------GY--- 248
                 4.46799*********54...........55678555555555555555...4555******999999.988888...55............67... PP

  GH79_allsf 239 tgtpedllntadtirlGsalildlglsyggss.lktisvHhYP 280
                 +  +e  +       ++d l+y+ s+ ++ +s+H+Y+
    Sequence 249 N--FEKQKQ-----------FVD--LAYAASYqVDFVSWHRYQ 276
                 8..877777...........777..57887778*********6 PP
```

Supplemental 2: Blast analysis to see the extent of the new family (GHxxx). Until proof of activity has been found for the more remote members, it is safer to restrict family membership to those sequences in the red box. In the meantime, sequences outside of the red box will be placed in the non-classified section of CAZy. We will reexamine the possible membership of these sequences to family GHxxx when family GHxxx has sufficient diversity (too few members right now)

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download ⌄ | GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
DUF5011 domain-containing protein [Paenibacillus sp. 32352]	806	806	100%	0.0	100.00%	WP_079915027.1
discoidin domain-containing protein [Paenibacillus sp. URHA0014]	728	728	100%	0.0	88.95%	WP_028553222.1
discoidin domain-containing protein [Paenibacillus whitsoniae]	727	727	100%	0.0	88.68%	WP_126144750.1
discoidin domain-containing protein [Paenibacillus sp. Soil766]	726	726	100%	0.0	88.68%	WP_057315858.1
discoidin domain-containing protein [Paenibacillus sp. Soil750]	724	724	99%	0.0	88.65%	WP_056617367.1
discoidin domain-containing protein [Paenibacillus oryzisoli]	722	722	100%	0.0	88.16%	WP_068663733.1
hypothetical protein BK126_18045 [Paenibacillus sp. FSL H7-0326]	583	583	100%	0.0	71.65%	OMC67485.1
discoidin domain-containing protein [Paenibacillus sp. FSL H7-0326]	583	583	100%	0.0	71.65%	WP_083655951.1
DUF5011 domain-containing protein [Paenibacillus sp. YN15]	582	582	98%	0.0	73.70%	WP_113018908.1
DUF5011 domain-containing protein [Paenibacillus sp. YN15]	224	224	90%	2e-61	38.27%	WP_113018909.1
discoidin domain-containing protein [Vallitalea guaymasensis]	121	121	94%	5e-26	27.30%	WP_113675418.1
hypothetical protein [Streptomyces griseus]	74.7	74.7	85%	1e-10	25.62%	WP_037672359.1
hypothetical protein [Micromonospora echinospora]	73.6	73.6	48%	3e-10	30.41%	WP_145834095.1
hypothetical protein [Paenibacillus aceris]	73.6	73.6	82%	4e-10	25.66%	NHW38917.1
hypothetical protein [Micromonospora echinospora]	72.8	72.8	48%	5e-10	30.41%	WP_088980648.1
hypothetical protein C6376_36625 [Streptomyces sp. P3]	72.8	72.8	85%	6e-10	25.62%	AVV46075.1
hypothetical protein [Streptomyces sp. P3]	72.4	72.4	85%	6e-10	25.62%	WP_159083366.1
hypothetical protein [Streptomyces sp. SLBN-192]	70.5	70.5	82%	3e-09	25.35%	WP_142165674.1
hypothetical protein GA0074692_4185 [Micromonospora pallida]	70.1	70.1	48%	4e-09	29.35%	SCL35799.1
carbohydrate-binding protein [Paenibacillus sp. YN15]	65.9	65.9	50%	1e-07	28.83%	WP_113018907.1
hypothetical protein ACM01_07090 [Streptomyces viridochromogenes]	64.7	64.7	82%	2e-07	23.71%	KMS75983.1
hypothetical protein [Streptomyces viridochromogenes]	64.7	64.7	82%	2e-07	23.71%	WP_159029107.1
hypothetical protein [Nocardioides sp. SLBN-172]	64.3	64.3	84%	3e-07	22.75%	WP_141790907.1