In silico identification of GH family 10 homologs in metagenomes

Jenny, Matt, Shannon, William

BIOS 700

April 30th, 2013

Goals!

1.) Practice the bioinformatics knowledge and computing skills by applying them to a research project, which might lead to novel research findings

2.) Learn how to design a bioinformatics workflow to answer biology/evolution questions

3.) Learn how to identify useful tools, datasets and existing knowledge from the research papers to help design the workflow

Let’s find out!

DID WE ACCOMPLISH THIS??

What are we looking at?

RESEARCH BACKGROUND

Glycoside hydrolases

- These enzymes hydrolyze the glycosidic bond between two (or more) carbohydrates, or between a carbohydrate and a non-carbohydrate moiety
- These enzymes were then broken down into different families based on sequence similarities

Mostly

But some are…

**This family also includes some cellobiohydrolases (EC:3.2.1.91)

Glycoside Hydrolase Family 10

Known Activities	endo-1,4-b-xylanase (EC 3.2.1.8); endo-1,3-b-xylanase (EC 3.2.1.32)
Mechanism	Retaining
Clan	GH-A
3D Structure Status	(b/a)8
Catalytic Nucleophile/Base	Glu (experimental)
Catalytic Proton Donor	Glu (experimental)
Note	formerly known as cellulase family F.
External resources	**CAZypedia**; HOMSTRAD; PRINTS; PROSITE;
Commercial Enzyme Provider(s)	MEGAZYME; NZYTech; PROZOMIX;
Statistics	GenBank accession (1849); Uniprot accession (932); PDB accession (120); 3D entries (25); cryst (1)

Summary | All (1516) | Archaea (8) | Bacteria (906) | Eukaryota (288) | unclassified (314) | Structure (25 - 1 cryst) | Characterized (260)

Predominant GH10 enzymes are denoted as EC 3.2.1.8, therefore classified as xylanases.

Table 1. Acronyms for genes and encoded enzymes

Enzyme	Gene	Protein	EC designations
Cellulase	*cel*	Cel	EC 3.2.1.4; EC 3.2.1.91
Xylanase	*xyn*	Xyn	EC 3.2.1.8
Mannanase	*man*	Man	EC 3.2.1.78
Lichenase	*lic*	Lic	EC 3.2.1.73; EC 3.2.1.58
Laminarinase	*lam*	Lam	EC 3.2.1.39

http://www.cazy.org/GH10.html

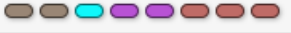
http://www.sciencedirect.com/science/article/pii/S0014579398002658

Domain organisations: Glycoside hydrolase, family 10 (IPR001000)

This family has the following domain organisations:

Showing **1** - **20** of **177**

1 2 3 4 5 6 7 8 9

Domain organisation	Number of proteins
	1777
	84
	77
	69
	66
	49
	46
	42
	41
	19
	18
	16
	15
	15
	11
	11
	10
	10
	7
	7

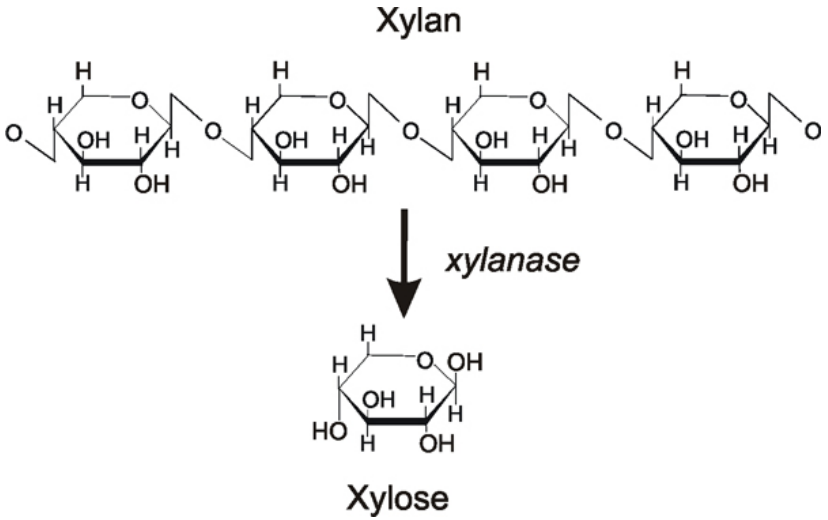
Colour key

- Carbohydrate-binding, CenC-like (IPR003305)
- S-layer homology domain (IPR001119)
- Cellulosome enzyme, dockerin type I (IPR016134)
- Ricin B lectin domain (IPR000772)
- Glycoside hydrolase, catalytic domain (IPR013781)
- Carbohydrate-binding domain, family 9 (IPR010502)
- Immunoglobulin-like fold (IPR013783)
- Galactose-binding domain-like (IPR008979)
- Cellulose-binding domain, fungal (IPR000254)
- Cellulose-binding domain, family II, bacterial-type (IPR001919)
- Glycosyl hydrolase, five-bladed beta-propellor domain (IPR023296)

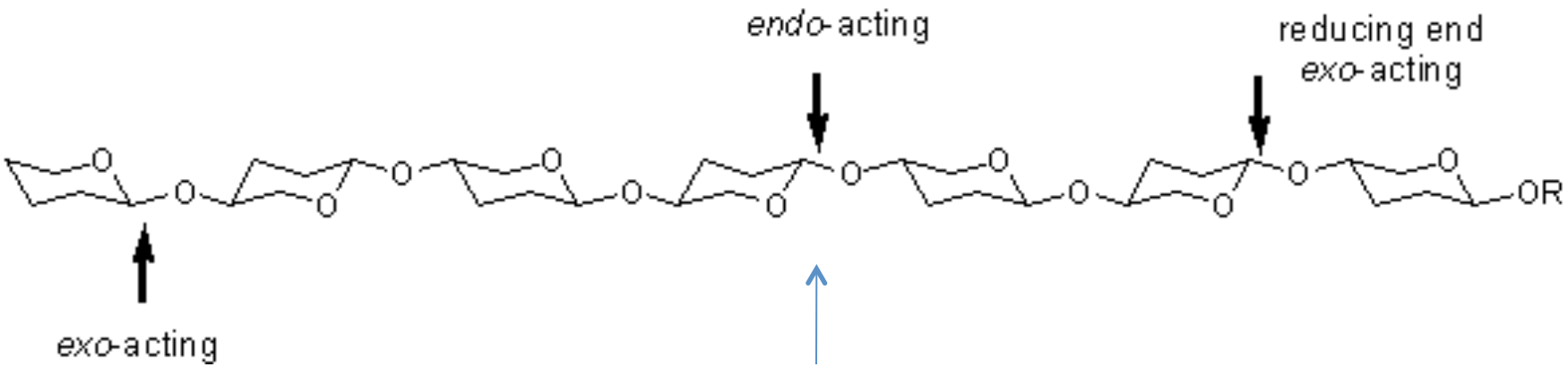
http://www.ebi.ac.uk/interpro/entry/IPR001000/domain-organisations

Endo-1,4-β-Xylanase

- Class of enzyme that degrades β-1,4-xylan into xylose
  - This breaks down one of the major component of plant cell walls, hemicellulose



http://www.biotek.com/assets/tech_resources/11087/figure3.jpg



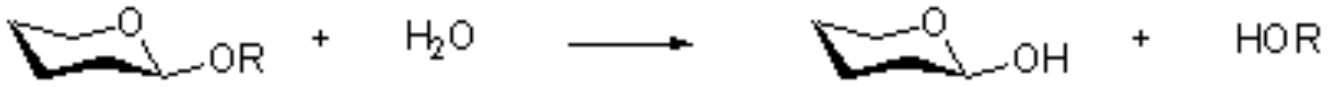
This enzyme cleaves β-1,4-xylan in an *endo* manner
(as seen above)

http://www.cazypedia.org/index.php/File:Retaining%26Inverting.png

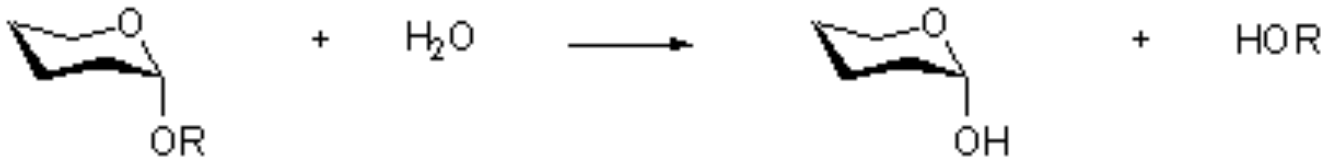
Retaining Mechanism

Retaining glycoside hydrolases:

1.) α

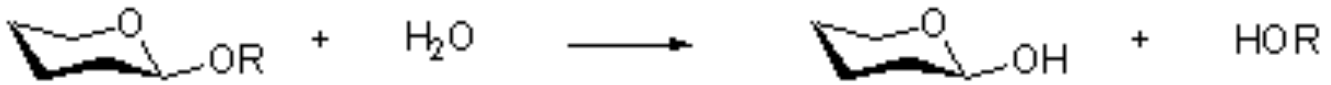


2.) β

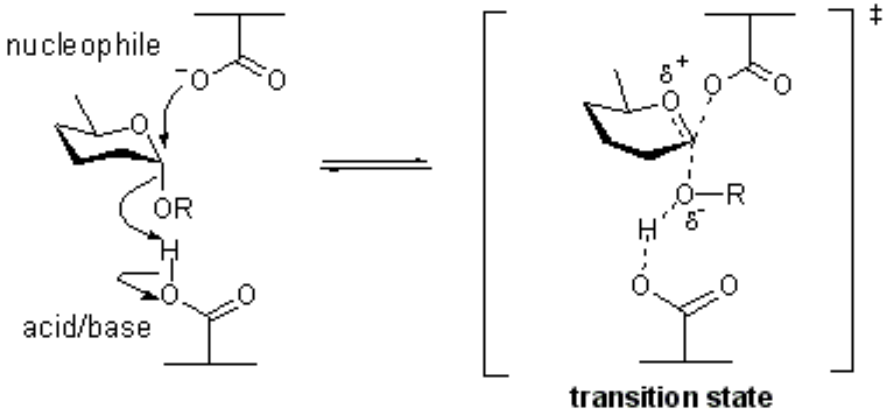


Retaining glycoside hydrolases:

1.) α

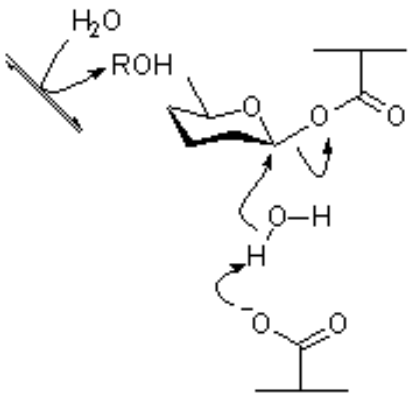


Retaining mechanism for an α-glycosidase:

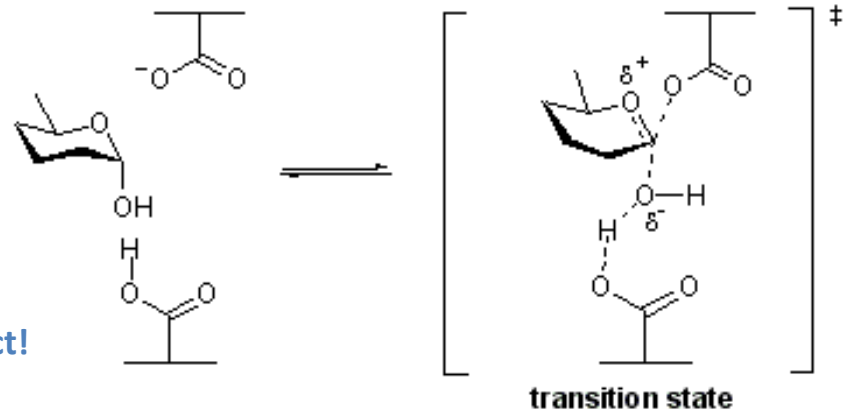


1.) Glycosylation step

2.) Oxocarbenium ion-like transition state



3.) Deglycosylation step



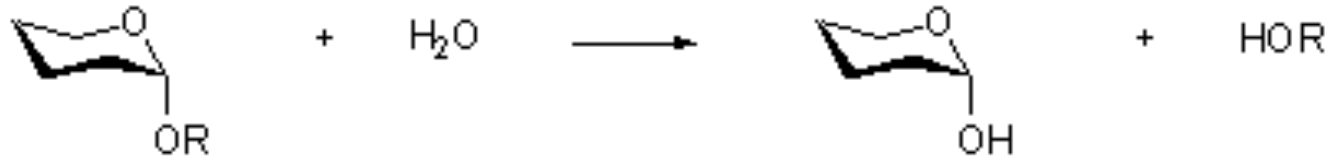
5.) Final product!

4.) Oxocarbenium ion-like transition state

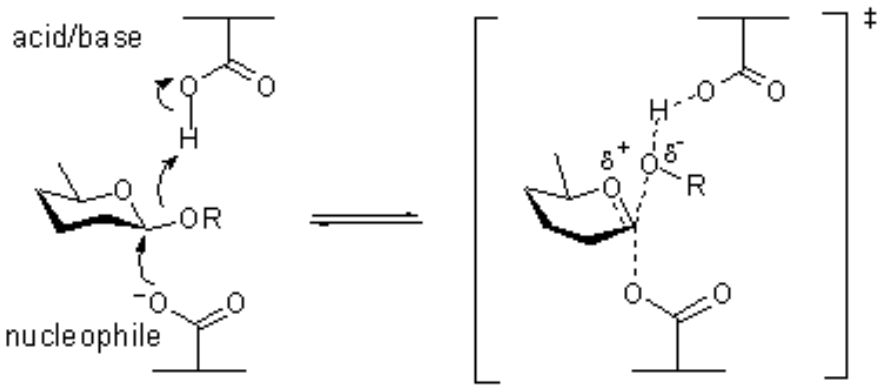
http://www.cazypedia.org/index.php/File:Retaining_glycosidase_mechanism_1.png

Retaining glycoside hydrolases:

2.) β

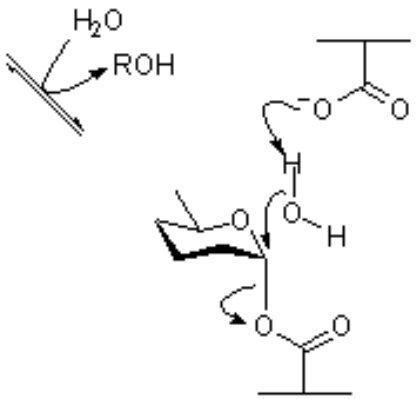


Retaining mechanism for a β-glycosidase:

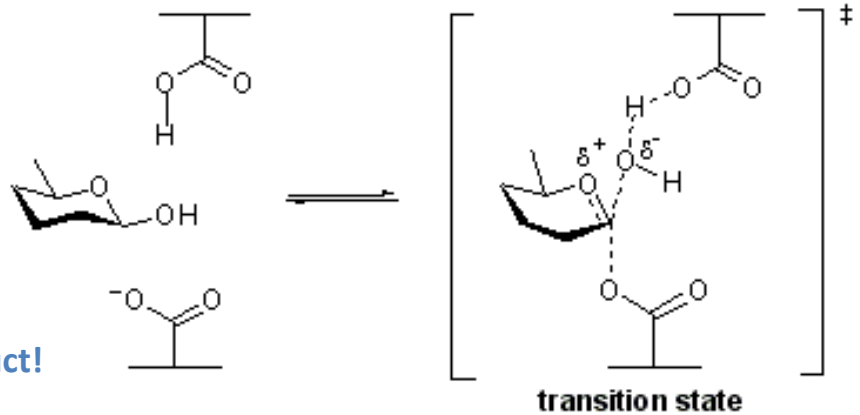


1.) Glycosylation step

2.) Oxocarbenium ion-like transition state



3.) Deglycosylation step



4.) Oxocarbenium ion-like transition state

5.) Final product!

http://www.cazypedia.org/index.php/File:Retaining_glycosidase_mechanism_1.png

3D Structure of GH10 / Active Sites

(the refined X-ray crystal structure at 1.8 Å, resolution of the catalytic core of the family F xylanase A from *Pseudomonas fluorescens* subspecies *cellulosa*)

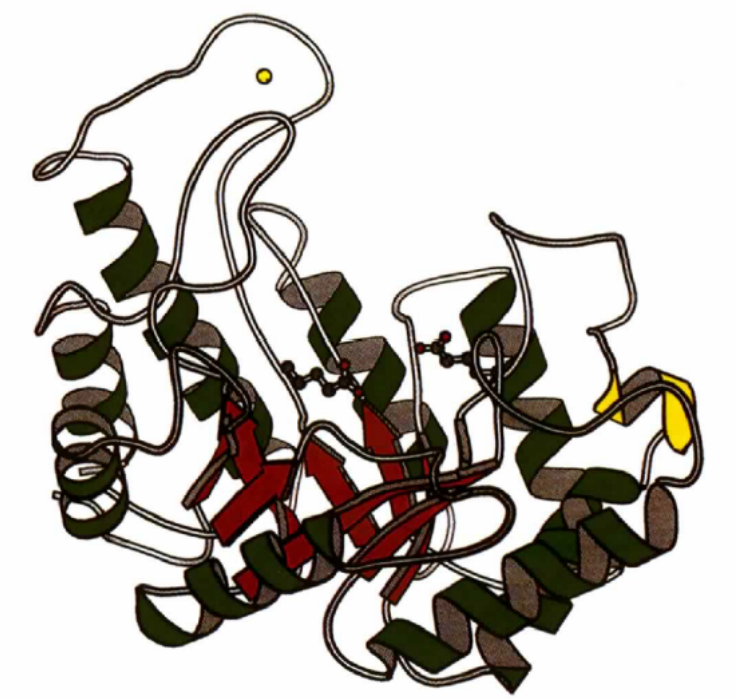
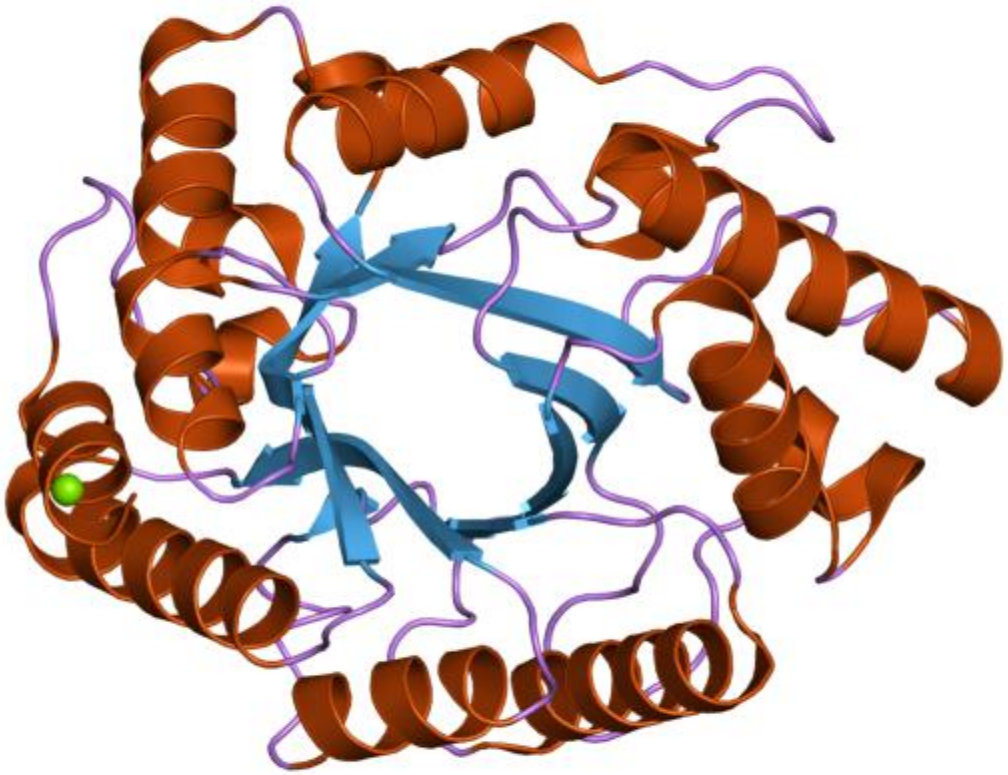


Fig. 1. The eightfold β/α-barrel of xylanase A viewed perpendicular to the barrel axis showing the acid base Glu127 on strand 4 and the nucleophile Glu246 on strand 7. The Ca atom (yellow dot) stabilizes the long loop after strand 7. This figure was prepared using *MOLSCRIPT* (Kraulis, 1991). The β-strands are shown as arrows in red and the eight α-helicies are shown in green; the additional α-helix ($\alpha 4a$) is shown in yellow. The glutamates are 5.5 Å apart consistent with a retention of the configuration during catalysis.

The active sites are the Glu127 on strand 4 and the Glu246 on strand 7

***These active sites are conserved across all homologous GH10 proteins**

Crystal structure of the catalytic domain of xylanase A from *Streptomyces halstedii* jm8



Again, note the eightfold β/α-barrel, a notable characteristic of GH10 family members

http://en.wikipedia.org/wiki/File:PDB_1nq6_EBI.jpg

Key Species

Key species	Number of proteins
Oryza sativa subsp. japonica (Rice)	33
Arabidopsis thaliana (Mouse-ear cress)	27

Taxa

- **cellular organisms** 2110 proteins | FASTA | Protein IDs
  - Archaea 18 proteins | FASTA | Protein IDs
    - Euryarchaeota 18 proteins | FASTA | Protein IDs
  - Bacteria (eubacteria) 1350 proteins | FASTA | Protein IDs
    - Actinobacteria 261 proteins | FASTA | Protein IDs
    - Bacteroidetes/Chlorobi group 245 proteins | FASTA | Protein IDs
    - Chlamydiae/Verrucomicrobia group 42 proteins | FASTA | Protein IDs
    - Chloroflexi 3 proteins | FASTA | Protein IDs
    - Cyanobacteria (blue-green algae) 33 proteins | FASTA | Protein IDs
    - Deinococcus-Thermus 3 proteins | FASTA | Protein IDs
    - Dictyoglomi 4 proteins | FASTA | Protein IDs
    - Fibrobacteres/Acidobacteria group 21 proteins | FASTA | Protein IDs
    - Firmicutes (Gram-positive bacteria) 339 proteins | FASTA | Protein IDs
    - Planctomycetes 23 proteins | FASTA | Protein IDs
    - Proteobacteria 248 proteins | FASTA | Protein IDs
    - Spirochaetes 16 proteins | FASTA | Protein IDs
    - Thermotogae 28 proteins | FASTA | Protein IDs
    - environmental samples 82 proteins | FASTA | Protein IDs
    - unclassified Bacteria 2 proteins | FASTA | Protein IDs
  - Eukaryota (eucaryotes) 742 proteins | FASTA | Protein IDs
    - Alveolata (alveolates) 8 proteins | FASTA | Protein IDs
    - Opisthokonta 402 proteins | FASTA | Protein IDs
    - Stramenopiles (heterokonts) 21 proteins | FASTA | Protein IDs
    - Viridiplantae 275 proteins | FASTA | Protein IDs
    - environmental samples 36 proteins | FASTA | Protein IDs
- **unclassified sequences** 561 proteins | FASTA | Protein IDs
  - environmental samples 556 proteins | FASTA | Protein IDs
  - metagenomes 5 proteins | FASTA | Protein IDs

http://www.ebi.ac.uk/interpro/entry/IPR001000/taxonomy

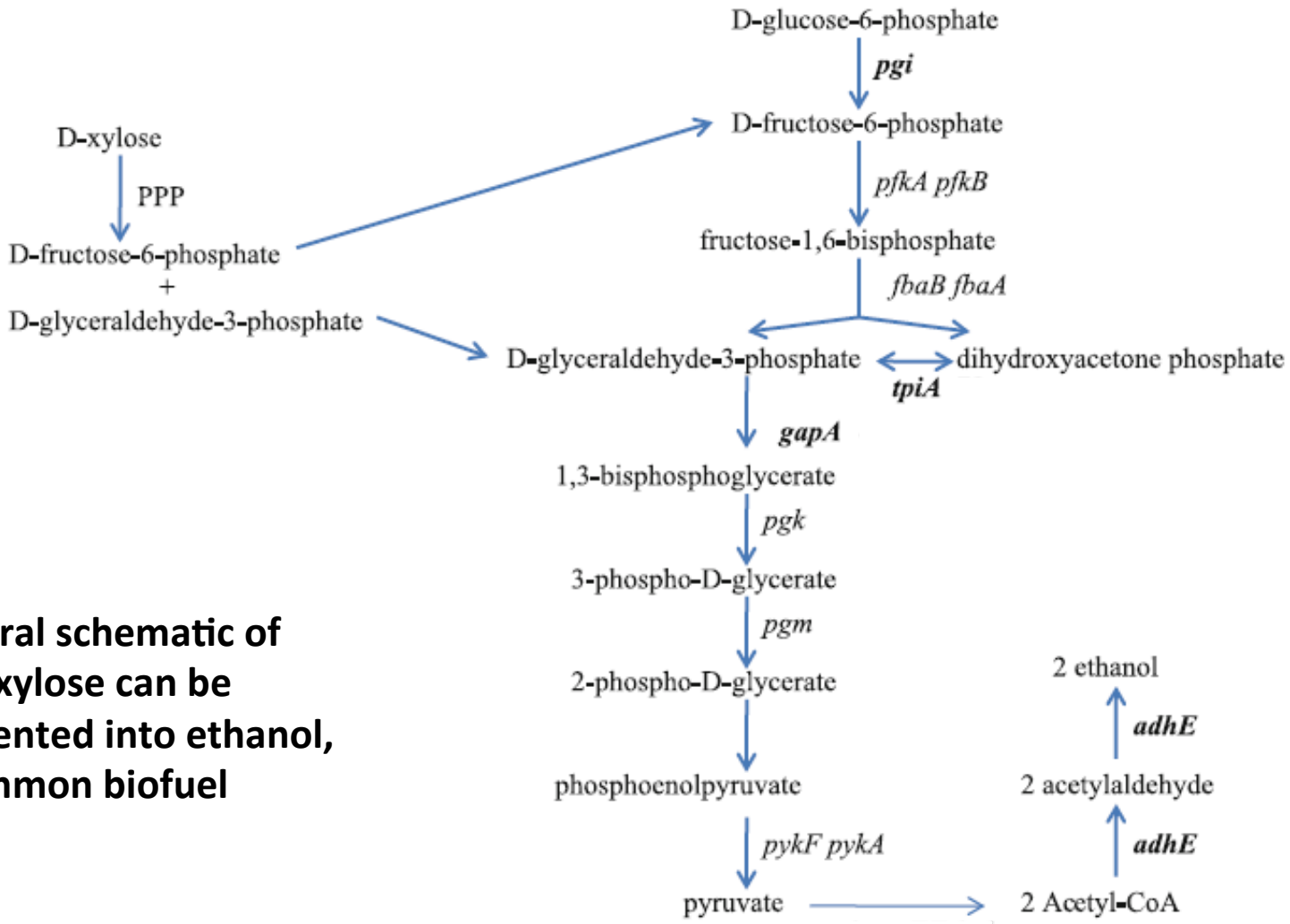
Roles of Xylanase

- Important role for microorganisms that live off of plant sources
  - An example of which is seen with fungi that degrade plant sources for nutrients
- Human uses include…
  - Chlorine-free bleaching of wood pulp prior to papermaking
  - Digesting silage (for cattle/biofuel)

Why is this important?

- Xylan is a major component of the hemicellulose portion of plant cell walls and constitutes up to 35% of the total dry weight of higher plants (Badal, 1999).
- Therefore studying xylanases helps contribute to bioenergy research and the creation of biofuels! ☺

General schematic of how xylose can be fermented into ethanol, a common biofuel



R. Manow, et. al., J Ind Microbiol Biotechnol ☺

What do we hope to accomplish?

GOALS

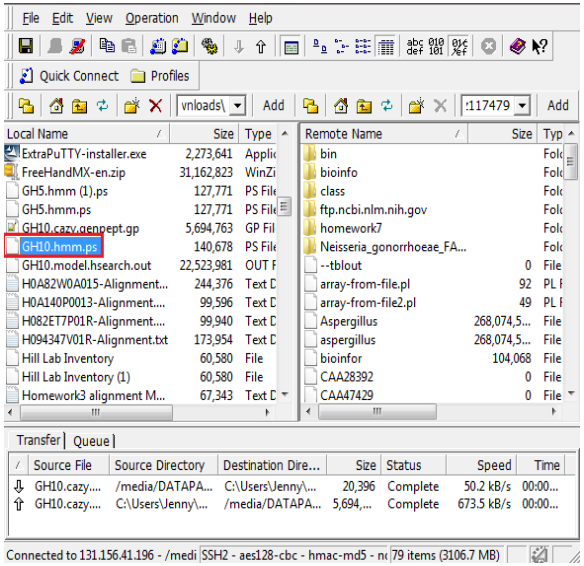
Compare the HMM domains of known GH10 proteins to determine if any JGI metagenome contain GH10 domains

How on earth did we do this?!

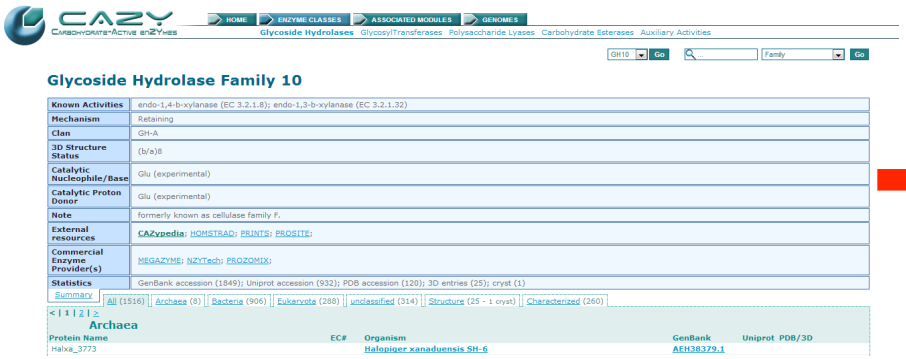
METHODS

Step 2 – Download Datasets

2.2.2: HMM model



2.2.1: Known GH10 proteins



1		Halopiger xanaduensis SH-6	AEH38379.1	
2		Halopiger xanaduensis SH-6	AEH38380.1	
3		Halorhabdus utahensis DSM 12940	ACV11171.1	C7NV87
4		Halorhabdus utahensis DSM 12940	ACV11291.1	C7NM83
5		Halorhabdus utahensis DSM 12940	ACV11653.1	C7NP66
6		Halorhabdus utahensis DSM 12940	ACV12558.1	C7NVN5
7		Haloterrigena turkmenica DSM 5511	ADB61053.1	D2RTV2
8		Natronococcus occultus SP4	AGB38895.1	
9		Acaryochloris marina MBIC11017	ABW28526.1	B0C274
10	3.2.1.8	Acidobacterium capsulatum 161	BAB40957.1	Q9AJR9
11		Acidobacterium capsulatum ATCC 51196	ACO33409.1	C1F8L6
12		Acidobacterium sp. MP5ACTX9	ADW71010.1	
13		Acidothermus cellulolyticus 11B ATCC 43068	ABK51955.1	A0LR95
14	3.2.1.8	Acidothermus cellulolyticus 11B ATCC 43068	ABK52146.1	A0LRT6
15		Acidovorax avenae subsp. citrulli AAC00-1	ABM34141.1	A1TT53
16	3.2.1.8	Actinomadura sp. FC7	AAA17888.1	Q59139

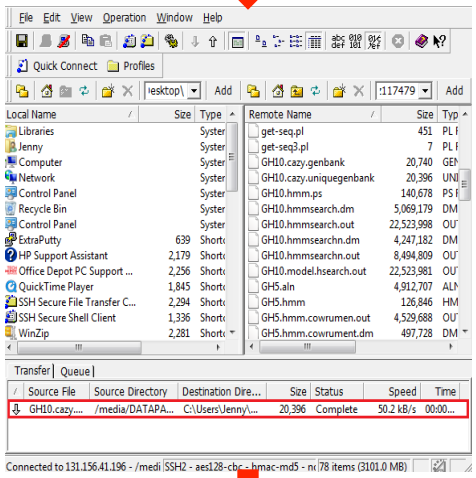
```
z117479@glu:~$ vi GH10.cazy.genbank
```

```
ACP87428.1
ACW02056.1
ACP87436.1
ACW02064.1
ACP87439.1
ACW02067.1
ACP87441.1
ACW02069.1
ACP87442.1
ACW02070.1
ACP87443.1
ACW02071.1
ACP87447.1
ACW02075.1
ACP87450.1
ACW02078.1
ACP87451.1
ACW02079.1
ACP87452.1
ACW02080.1
ACP87455.1
ACW02083.1
ACP87458.1
-- INSERT --
```

Step 2.2.1 – Known GH10 Proteins

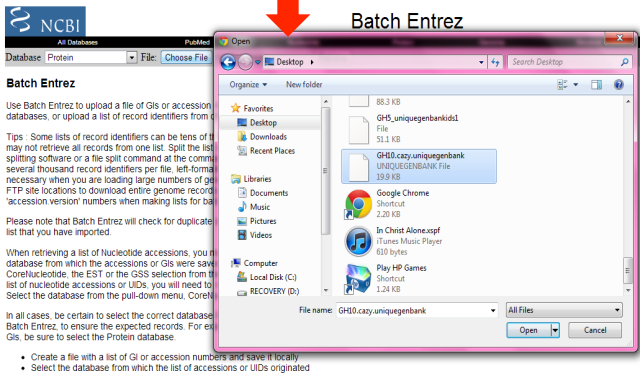
Parse for unique GenBank ID's

```
z117479@glu:~$ less GH10.cazy.genbank | sed '/^$/d' | sort | uniq | sort > GH10.
cazy.uniquegenbank
z117479@glu:~$ head GH10.cazy.uniquegenbank
A43802
AAA16427.1
AAA17888.1
AAA21812.1
AAA23059.1
AAA23062.1
AAA23227.1
AAA23286.1
AAA56791.1
AAA56792.1
```



Transfer unique ID file to local computer

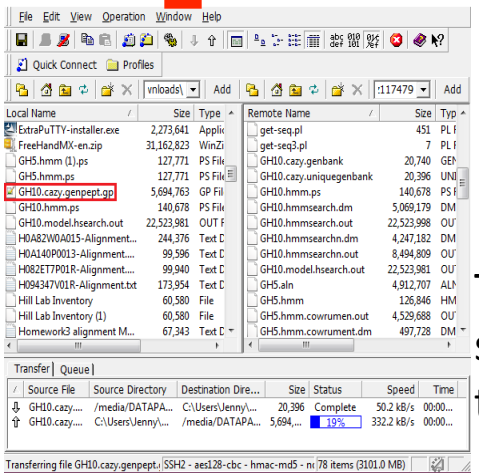
Batch Entrez



Search database for sequences

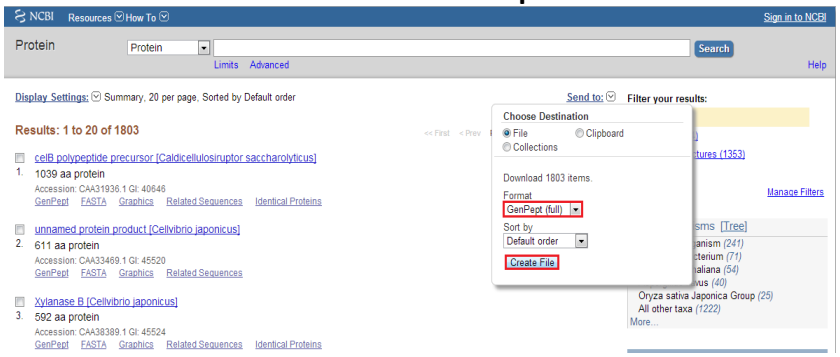
Reformat GenBank sequences to FASTA sequences

```
z117479@glu:~$ seqret -sequence GH10.cazy.genpept.gp -outseq GH10.cazy.fasta -sf
ormat genbank -osformat fasta
```



Transfer GenBank sequences to terminal

Download GenBank sequence to local computer



Step 3-4.1.1: Homology Search and Parsing Output Files

Hmmrsearch and parsing of JGI metagenome proteins against GH10 HMM

```
z117479@glu:~/project1$ hmmsearch --domtblout GH10.hmm.jgi.dm /media/DATAPART3/z
117479/GH10.hmm.ps /home/yyin/db/jgi-v3.5.pr.fa > GH10.hmm.jgi.dm &
z117479@glu:~/project1$ less GH10.hmm.jgi2.dm | grep -v '^#' | awk '{print $1,$3
,$6,$7,$12,$13,$16,$17,$18,$19}' | awk '$6<1e-5&&($8-$7)/$3>.8' | sed 's/ /\t/g'
 > GH10.hmm.jgi.dm.ps
```

Hmmrsearch and parsing of GH10 known NCBI-nr proteins against GH10 HMM model

```
z117479@glu:~$ hmmsearch --noali --domtblout GH10.hmm.NCBI-nr.dm GH10.hmm.ps GH1
0.cazy.fasta > GH10.hmm.NCBI-nr.out &
z117479@glu:~/project1$ less GH10.hmm.NCBI-nr.dm | grep -v '^#' | awk '{print $1
,$3,$6,$7,$12,$13,$16,$17,$18,$19}' | awk '$6<1e-5&&($8-$7)/$3>.8' | sed 's/ /\t
/g' > GH10.hmm.NCBI-nr.dm.ps
z117479@glu:~/project1$ less GH10.hmm.NCBI-nr.dm.ps
```

```
CCA71122        1168    303     0       8.5e-102        8.5e-102        5
301     126     429
CCA71122        1168    303     0       1.1e-100        1.1e-100        5
301     474     777
CCA71122        1168    303     0       3.5e-101        3.5e-101        5
301     822     1125
ADG76430        820     303     1.8e-214        1.2e-107        1.2e-107
4       302     44      338
ADG76430        820     303     1.8e-214        1e-107  1e-107  5       302
402     696
ADG75375        815     303     1.3e-205        4.1e-102        4.1e-102
4       302     42      337
ADG75375        815     303     1.3e-205        5.4e-104        5.4e-104
3       303     395     691
ABX44209        2457    303     1.9e-154        8.1e-85 8.2e-85 2       302
391     722
ABX44209        2457    303     1.9e-154        2.1e-68 2.1e-68 2       302
744     1083
ADB36415        366     303     3.7e-122        4.2e-122        4.2e-122
2       302     37      365
CCH02308        352     303     1.9e-120        2.2e-120        2.2e-120
2       302     22      351
AEE95681        380     303     6.9e-120        7.9e-120        8e-120  1
GH10.hmm.NCBI-nr.dm.ps
```

Step 4.2-4.3: Parse Files to Extract Protein Sequences

Extract hit ID's

```
z117479@glu:~/project1$ less GH10.hmm.jgi.dm.ps | cut -f1 > GH10.hmm.jgi.dm.ps.i
d
z117479@glu:~/project1$ less GH10.hmm.NCBI-nr.dm | grep -v '^#' | awk '{print $1
,$3,$6,$7,$12,$13,$16,$17,$18,$19}' | awk '$6<1e-5&&($8-$7)/$3>.8' | sed 's/ /\t
/g' | cut -f1 > GH10.hmm.NCBI-nr.dm.id
```

Parse for unique ID's

```
z117479@glu:~/project1$ less GH10.hmm.jgi.dm.ps.id | sort | uniq | sort > GH10.h
mm.jgi.dm.ps.uniqid
z117479@glu:~/project1$ less GH10.hmm.NCBI-nr.dm.id | sort | uniq | sort > GH10.
hmm.NCBI-nr.dm.uniqid
```

```
CCA71122
CCA71122
CCA71122
ADG76430
ADG76430
ADG75375
ADG75375
ABX44209
ABX44209
ADB36415
CCH02308
AEE95681
ACP87343
ACW01971
ACX75889
ACX75889
ADL24782
ADL24782
AEV97130
ACP87323
ACW01951
AEE49315
BAM46412
GH10.hmm.NCBI-nr.dm.id
```

Create perl script using bioperl modules to extract sequences from hit ID's

```
z117479@glu:~/project1$ vi step4.3.pl
```

```
#!/usr/bin/perl -w

open (ID,$ARGV[0]);
while (<ID>) {
        chomp $_;
        $id_hash{$_}=1;
}

use Bio::SeqIO;

$new=Bio::SeqIO->new(-file=>$ARGV[1], -format=>"fasta");

while ($seq=$new->next_seq){
        if (defined $id_hash{$seq->id}){
        print ">",$seq->id,"\n",$seq->seq,"\n";
        }
}
```

Step 4.3 – Extract Protein Sequences

Run perl script on parsed hit ID files and original databases, JGI metagenome and known GH10 proteins (NCBI)

```
z117479@glu:~/project1$ nohup perl step4.3.pl GH10.hmm.jgi.dm.ps.uniqid /home/yy
in/db/jgi-v3.5.pr.fa > GH10.hmm.jgi.id.totalseq &
[1] 50501
z117479@glu:~/project1$ nohup perl step4.3.pl GH10.hmm.NCBI-nr.dm.uniqid GH10.ca
zy.fasta > GH10.hmm.NCBI-nr.id.totalseq &
[2] 50854
>2000622610
MHKRLLPLVLALVLVVTPFLLANPATPAAAAEIPWGDKSTPFGMVVSLGNRVRSDEMPTMIRLMREAGVQWNREEIWWDQ
VQFEPDGPFRWDGDSSRFYNYDRAIQLQAEAGISILGLLDYNPAWFKGRNPHPEEWLSDWGDYVYATVARYGRDRGQIKY
WEVWNEPNLVPSGYESGLYNVDDFVRVLQTASAAIRAADPEAKIVLGGVADIWSEIPEFAYDTPDYLQRLYALDAWPMFD
ILGLHPYRPDAPEVPVLRRDRSQTYREQAAEIDALLAQFGNKPIWYTEVGWSTESDGIVSEDEQAAWLQRFYLLAMTHPG
VEKIFWYDFRNDTGDNSNYTRPINDPNENQFHFGLLRGNYPLNFDDQRMRKPSYSAYYHLNHNLGGVSWRAKYALP
>2001264806
MAFSKDKASFTRRSAIAAGLAAGVSACVTPQPGVPAMPSLDQLAKRSGRRFGSAVGWGRPGADRGSFANPAYAAILENEC
SLLVPENELKWQWTRPGPGQFDFRQFDAIADYASRKGLLLRGHTLFWLPEKWYPKWLVAHDFGAQPAAKAEAMLREHVQT
VCRRYGTRIYSYDVVNEAVQPENGLIRDTVVTKALGHEATLDVMFQTARAEAPHAQLVYNDYMSWERNSEDETHMRGVLK
LLEGFRKRGTPVDALGIQSHIRLLKPLTVAEIVRESGGPWRRFLDEVVAMGYKLEITEFDVNDRLAPTNPLERDRMVADY
AKAYLDVMLSYPQLRDIMAWGMVDRYSWLTGFDPRKDKTLKRGTPYDKHFRPKLLREAIAAAFMGAAAPA
>2003420141
MTPTRRAILAAPLALAACDRFASAEPSPPNVPPLKSIAPAPFGTAIKASQIDDPDWVALARANVSQLTPEWEMKMEYILA
DGLDRPNFDRSDRIAAFARAEGMAMHGHTLIWYAQGKEAFAGLSGAAFDRAFDGYIATVAGRYRGKVRSWDVVNEPILDD
GSGMRDCHWSARYGHDGYILRAFEKARIADPDAVLFLNEYNQESVPAKGAQFLKLVERLLKAGCPLQGVGLQSHLWIDIP
EGVIAAYMREISQFGLPIHVSELDCTLRTENRLDLRSQADRIAAQGARVTELASAFAALPKAQQFAFTVWGLRDTD
>2003425276
LGIGAQALDGMTTVLHIKVQDFPTHAANLPVAMIPNCAATRHAHFTLDGGCPRRHGGADAANAAGPAAPGCASPRSPLTM
WFTRVSLANPVFATMLMLALMVLGVFSYQRLKVDQFPNVDFPVVVVTVDYPGASPEIVESEVTKKIEEGVNSIAGINALT
SRSYESTSVVIIEFQLHVDGRRAAEDVREKVASVRPTLRDEVKEPRVLRFDPASSPIWSVAVLPQADAGAKAPDAVALTS
WADQVLKKRLENVRGVGAVNLVGATPREINIYLQPEALEAYAITPDVPPLKSIAPAPFGTAIKAMQIDDPDWVALARANV
SQLTPEWEMKMEYILANGLDRPNFDRTDRIAAFAQAQGMAMHGHTLIWYSQGQEAFAGMDDAAFDRAFDGYIAAVAGRYR
GH10.hmm.jgi.id.totalseq
```

Step 5: Retrieve HMM Domain Regions

Perl script using bioperl modules to extract HMM domain regions using first HMM parsed output files

```
z117479@glu:~/project1$ vi idsubseq-bioperl.pl
#!/usr/bin/perl

use Bio::SeqIO

open (IN,$ARGV[0]);

while (<IN>){
        chomp$_;
        @col=split(/\t/,$_);
        push(@{$id_hash{$col[0]}},$_);
}

$new=Bio::SeqIO->new(-file=>$ARGV[1], -format=>"fasta");

while($seq=$new->next_seq){
        if(defined $id_hash{$seq->id}){
        @id_array=@{$id_hash{$seq->id}};

        foreach(@id_array){
        @id_col=split(/\t/,$_);
        print ">",$id_col[0],"|",$id_col[-2],"-",$id_col[-1],"\n",
        $seq->subseq($id_col[-2],$id_col[-1]),"\n";
        }
        }
}
```

```
>2000622610|77-378
WWDQVQFEPDGPFRWDGDSSRFYNYDRAIQLQAEAGISILGLLDYNPAWFKGRNPHPEEWLSDWGDYVYATVARYGRDRG
QIKYWEVWNEPNLVPSGYESGLYNVDDFVRVLQTASAAIRAADPEAKIVLGGVADIWSEIPEFAYDTPDYLQRLYALDAW
PMFDILGLHPYRPDAPEVPVLRRDRSQTYREQAAEIDALLAQFGNKPIWYTEVGWSTESDGIVSEDEQAAWLQRFYLLAM
THPGVEKIFWYDFRNDTGDNSNYTRPINDPNENQFHFGLLRGNYPLNFDDQRMRKPSYSAYY
>2001264806|42-381
QLAKRSGRRFGSAVGWGRPGADRGSFANPAYAAILENECSLLVPENELKWQWTRPGPGQFDFRQFDAIADYASRKGLLLR
GHTLFWLPEKWYPKWLVAHDFGAQPAAKAEAMLREHVQTVCRRYGTRIYSYDVVNEAVQPENGLIRDTVVTKALGHEATL
DVMFQTARAEAPHAQLVYNDYMSWERNSEDETHMRGVLKLLEGFRKRGTPVDALGIQSHIRLLKPLTVAEIVRESGGPWR
RFLDEVVAMGYKLEITEFDVNDRLAPTNPLERDRMVADYAKAYLDVMLSYPQLRDIMAWGMVDRYSWLTGFDPRKDKTLK
RGTPYDKHFRPKLLREAIAA
>2003420141|40-315
APFGTAIKASQIDDPDWVALARANVSQLTPEWEMKMEYILADGLDRPNFDRSDRIAAFARAEGMAMHGHTLIWYAQGKEA
FAGLSGAAFDRAFDGYIATVAGRYRGKVRSWDVVNEPILDDGSGMRDCHWSARYGHDGYILRAFEKARIADPDAVLFLNE
YNQESVPAKGAQFLKLVERLLKAGCPLQGVGLQSHLWIDIPEGVIAAYMREISQFGLPIHVSELDCTLRTENRLDLRSQA
DRIAAQGARVTELASAFAALPKAQQFAFTVWGLRDT
>2003425276|296-603
APFGTAIKAMQIDDPDWVALARANVSQLTPEWEMKMEYILANGLDRPNFDRTDRIAAFAQAQGMAMHGHTLIWYSQGQEA
FAGMDDAAFDRAFDGYIAAVAGRYRGKVRSWDVVNEPILDDGSAMRDCHWSARYGQDGYILRAFEKARIADPEAVLFLNE
YNQESVPAKGVQFLKLVERLLKAGCPLQGLGLQSHLWIDIPEGVIAAYLREVAQFGLPIHVSELDCTLRTENRLDMRNKA
DRIAAQGARVTELASAFAALPKAQQFAFTAWGLRDTDSWYRQGEKDDGKDKPLPFDSFGRPNPMAAAL
>2003454879|40-355
LAPFPVGATGMTGQLDDPTWVAAVDRHLDQITPEWELKPERILGPGFAYDFSAADRMVDWATARGLRVHGHALVWYAQGL
GH10.hmm.jgi.id.domseq
```

Run perl script using the sequences extracted from the hit ID's (step 4.3) and the first parsed hmmrsearch parsed output files (step 4.1.1)

```
qz117479@glu:~/project1$ nohup perl idsubseq-bioperl.pl GH10.hmm.jgi.dm.ps GH10.
mm.jgi.id.totalseq > GH10.hmm.jgi.id.domseq &
[4] 53635
z117479@glu:~/project1$ nohup perl idsubseq-bioperl.pl GH10.hmm.NCBI-nr.dm.ps GH
10.hmm.NCBI-nr.id.totalseq > GH10.hmm.NCBI-nr.id.domseq &
```

Step 6: Combine JGI and NCBI files

Use “cat” to combine the JGI metagenome and known GH10 proteins (NCBI) files (step 5) containing the HMM domain sequence regions into one new file

```
z117479@glu:~/project1$ cat GH10.hmm.jgi.id.domseq GH10.hmm.NCBI-nr.id.domseq >
GH10.hmm.jgi.NCBI-nr.domseq
z117479@glu:~/project1$ less GH10.hmm.jgi.NCBI-nr.domseq
>2031944243|75-350
RGEYEPHEGETQAHRIRATAEWFAARAVRLKGHPLVWHTVKAPWMDSLSPDAAWHNTLDRIRREVGDFSGLIDTWDVVNE
AVIMPVFVNEPDGTPNVISRIAQQKGRIPLIREAFDEARATNPGATLLLNDFDLSSAYEILIEGLLEAGVRIDALGLQTH
MHQGYWGEETMLAMVDRFARYGLPLHLTENTLLSGDLMPAHIVDLNDYQPASWPSTPEGEARQADEIARHYRSLVAHPAV
ASITYWGISDAGAWLGAPAGLLRADGSRKPAYDALR
>CAA31936|45-372
lcesykddfmigvaiparclsndtdkrmvlkhfnsitaenemkpesllagqtstglsyrfstadafvdfastnkigirgh
tlvwhnqtpdwffkdsngqrlskdallarlkqyiydvvgrykgkvyawdvvneaidenqpdsyrrstwyeicgpeyieka
fiwaheadpnaklfyndynteiskkrdfiynmvknlkskgipihgigmqchinvnwpsvseiensiklfssipgieihit
eldmslynygssenystppqdllqkqsqkykeiftmlkkyknvvksvtfwglkddyswlrsfygkndwpllffedysakp
aywaviea
>CAA33469|270-607
adfpigvavaasggnadiftssarqnivraefnqitaenimkmsymysgsnfsftnsdrlvswaaqnggtvhghalvwhp
syqlpnwasdsnanfrqdfarhidtvaahfagqvkswdvvnealfdsaddpdgrgsangyrqsvfyrqfggpeyideafr
rapradptaelyyndfnteengakttalvnlvqrllnngvpidgvgfqmhvmndypsianirqamqkivalsptlkikit
eldvrlnnpydgnssndytnrndcavscagldrqkarykeivqaylevvppgrrggitvwgiadpdswlythqnlpdwpl
lfndnlqpkpayqgvvea
>CAA38389|321-562
akkfignittsgavrsdftrywnqitpeneskwgsvegtrnvynwapldriyayarqnnipvkahtfvwgaqspswlnnl
sgpevaveieqwirdycarypdtamidvvneavpghqpagyaqrafgnnwiqrvfqlarqycpnsililndynnirwqhn
efialakaqgnyidavglqahelkgmtaaqvktaidniwnqvgkpiyiseydigdtndqvqlqnfqahfpvfynhphvhg
it
```

Step 8: Build Multiple Sequence Alignment

Use MAFFT program to build multiple sequence alignment using the newly combined file (step 6)

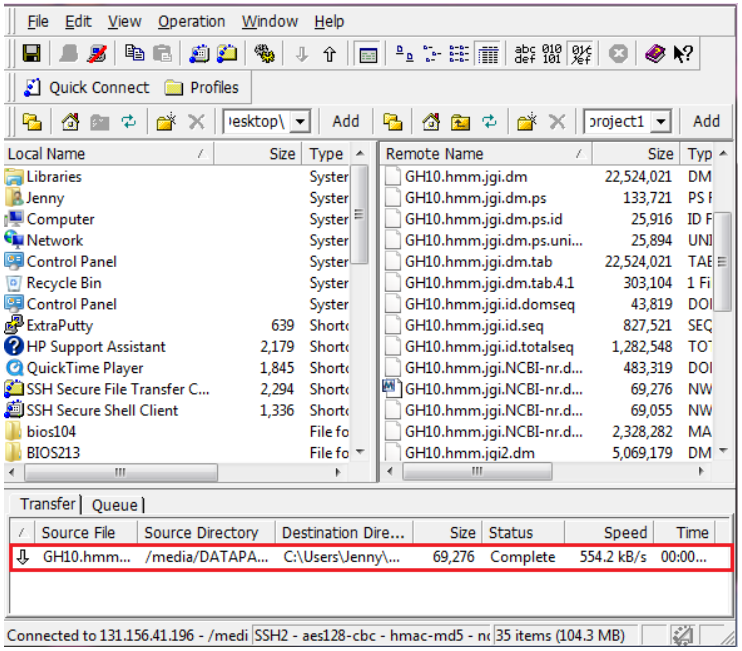
```
z117479@glu:~/project1$ mafft --anysymbol GH10.hmm.jgi.NCBI-nr.domseq > GH10.hmm
.jgi.NCBI-nr.domseq.mafft &
```

```
>2000622610|77-378
------------------------------------------------------------
-------------------------WWD-QVQF-EPDGPFRW-------------------
---------------------------------DGD--------------------SSRFY
NY----------------------DRA---------IQLQ---A-EAGI---S------I
LG--L-----------------LDY------NPAWFKGRNPH-----------------P
EE-W-L------------------------------------------------------
--------------------SDW-GDY-------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
--------------V-YATVA-----R-----Y----------GRDRG-----Q-------
----------------------IKYW----------------EVW-NEPNLVP-------
------------------------------------------------------------
------------------------------------------------------------
--------SGYESGLYNVDDFVRVLQTASAAIRAADPEAKIV-------------------
LGG-V---------ADI------------WSE--------IPEFAY-DTP----------
--------DYLQRLYAL--------------------D------------------AWPM
FD--ILGLHPYRPDAPEVPVLRRDRSQTYREQAAE--------------------------
----------------------IDALLAQF--G--------N-----------------
--KPIWYTEVG---------------WST----ESDGI-----------VSEDE------
GH10.hmm.jgi.NCBI-nr.domseq.mafft
```

Step 8: Build Phylogenetic Tree

Use FastTree to build phylogenetic tree from the MAFFT alignment

```
z117479@glu:~/project1$ nohup /home/mrupani/Downloads/FastTree GH10.hmm.jgi.NCBI
-nr.domseq.mafft > GH10.hmm.jgi.NCBI-nr.domseq.fasttree.nwk &
```

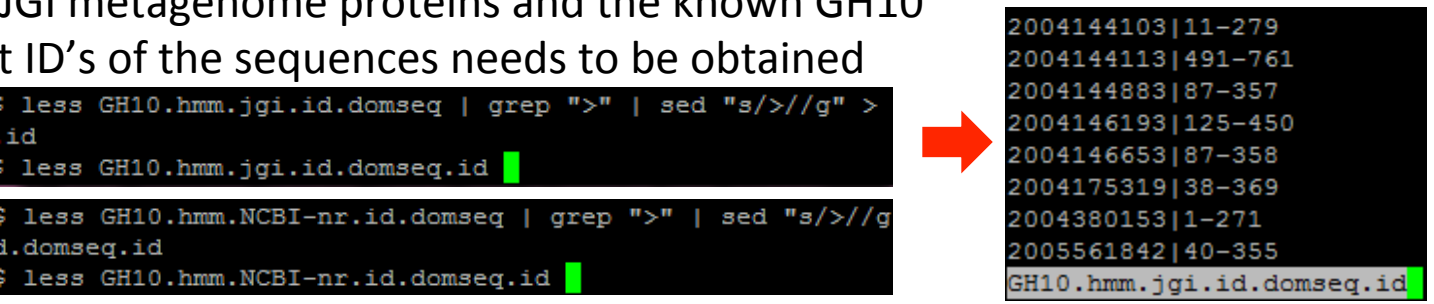


Step 9: Parse HMM Domain Sequence Files for hit ID's

To color code the JGI metagenome proteins and the known GH10 proteins, the exact ID's of the sequences needs to be obtained

```
z117479@glu:~/project1$ less GH10.hmm.jgi.id.domseq | grep ">" | sed "s/>//g" > GH10.hmm.jgi.id.domseq.id
z117479@glu:~/project1$ less GH10.hmm.jgi.id.domseq.id
```

```
z117479@glu:~/project1$ less GH10.hmm.NCBI-nr.id.domseq | grep ">" | sed "s/>//g" > GH10.hmm.NCBI-nr.id.domseq.id
z117479@glu:~/project1$ less GH10.hmm.NCBI-nr.id.domseq.id
```



Perl scripts were written to assign the hex blue color code to JGI metagenome proteins and the hex red color code to known GH10 proteins

```
z117479@glu:~/project1$ vi jgicolor.pl
#!/usr/bin/perl

while (<>){
        chomp $_;
        print $_, "\t#0000FF\n";
}
```

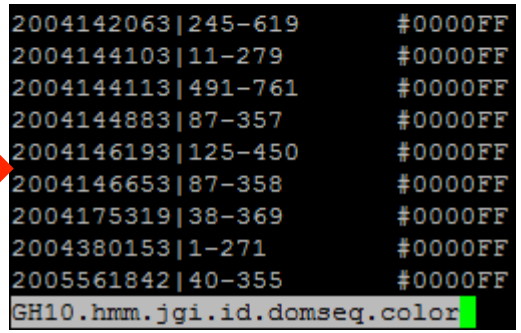
```
z117479@glu:~/project1$ vi NCBI-nrcolor.pl
#!/usr/bin/perl

while(<>){
        chomp$_;
        print $_, "\t#FF0000\n";
}
```

Perl script was run to assign the color codes to the previously parsed files

```
z117479@glu:~/project1$ perl jgicolor.pl GH10.hmm.jgi.id.domseq.id > GH10.hmm.jgi.id.domseq.color
z117479@glu:~/project1$ less GH10.hmm.jgi.id.domseq.color
```

```
z117479@glu:~/project1$ perl NCBI-nrcolor.pl GH10.hmm.NCBI-nr.id.domseq.id > GH10.hmm.NCBI-nr.id.domseq.color
z117479@glu:~/project1$ less GH10.hmm.NCBI-nr.id.domseq.color
```



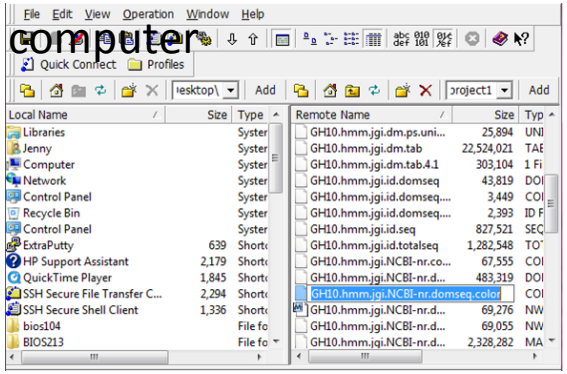
Step 9: Download Datasets to computer

Combine color files

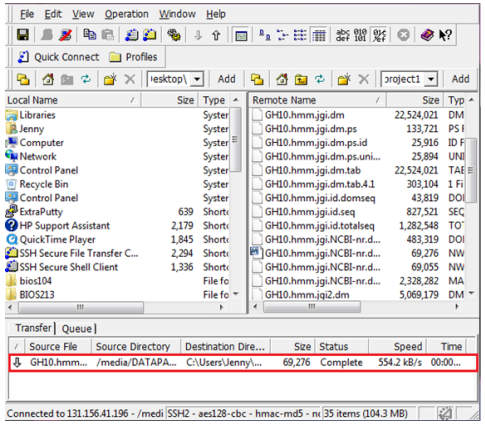
```
z117479@glu:~/project1$ cat GH10.hmm.jgi.id.domseq.color GH10.hmm.NCBI-nr.id.dom
seq.color > GH10.hmm.jgi.NCBI-nr.domseq.color
z117479@glu:~/project1$ less GH10.hmm.jgi.NCBI-nr.domseq.color
```

```
2029678161|71-373        #0000FF
2029683389|3-296         #0000FF
2029724659|1-347         #0000FF
2030993216|7-327         #0000FF
2031944243|75-350        #0000FF
CAA31936|45-372 #FF0000
CAA33469|270-607         #FF0000
CAA38389|321-562         #FF0000
CAA43712|3-337  #FF0000
AAA23059|17-340 #FF0000
AAA23062|18-300 #FF0000
```

Transfer combined color files to local computer

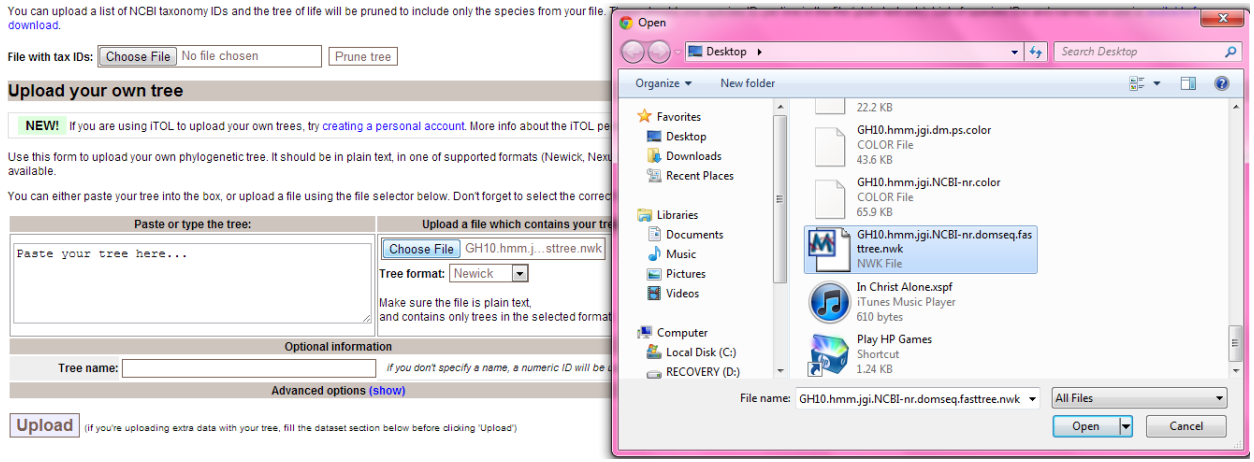


Transfer FastTree Newick file to local computer

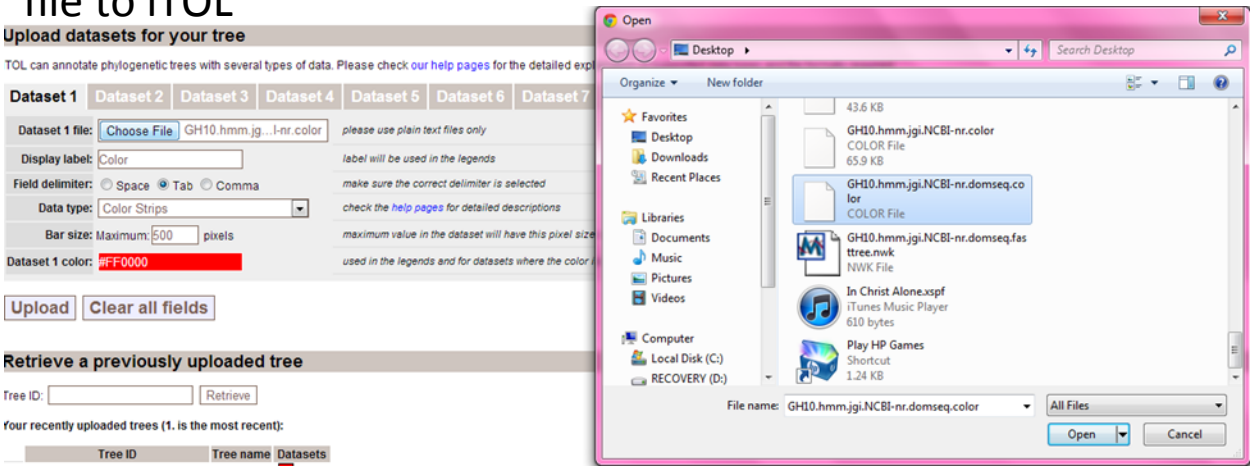


Step 9: Upload Datasets to iTOL

Upload JGI metagenome and known GH10 protein Newick file to iTOL

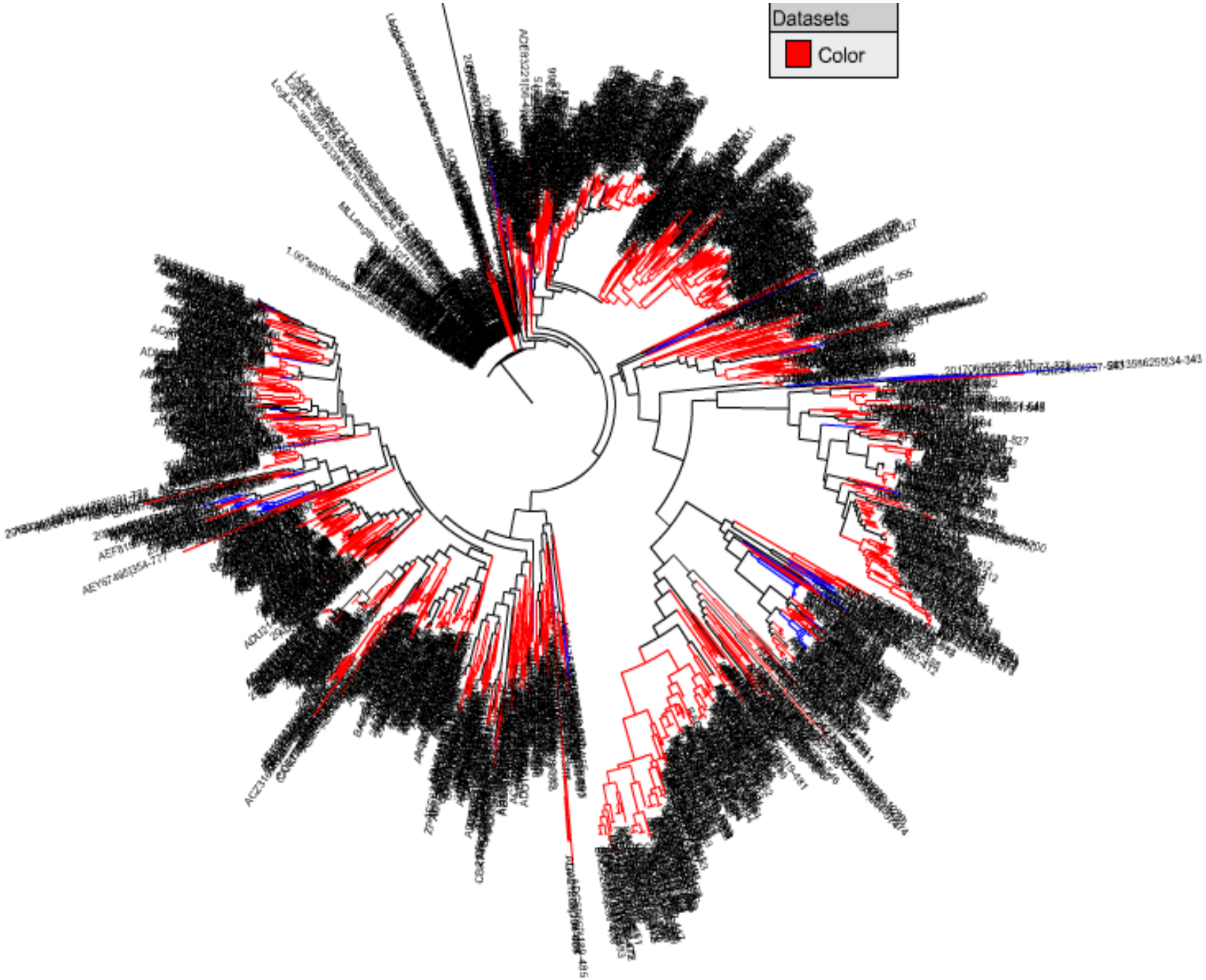


Upload JGI metagenome and GH10 protein combined color file to iTOL

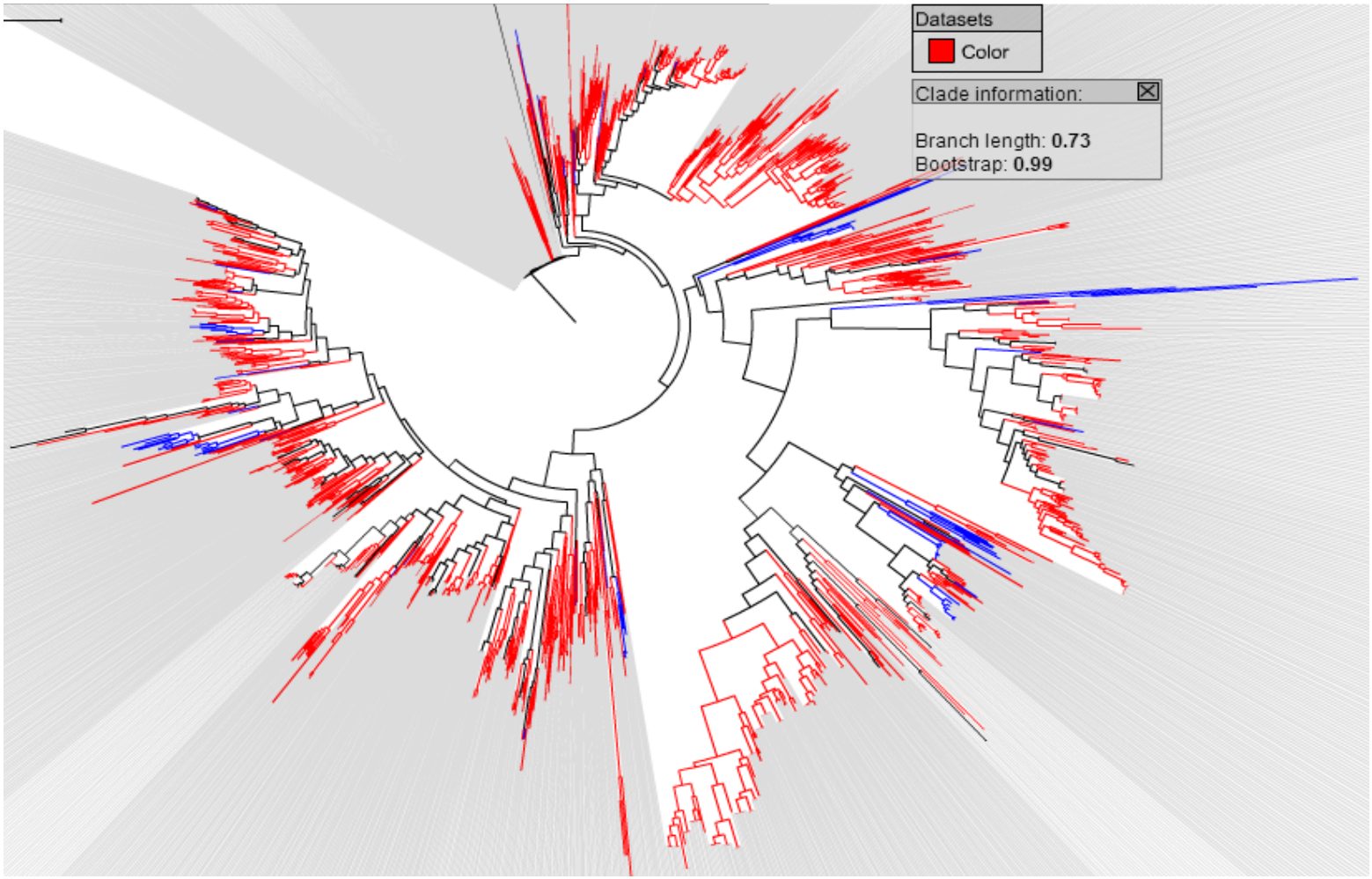


The fun stuff… ☺

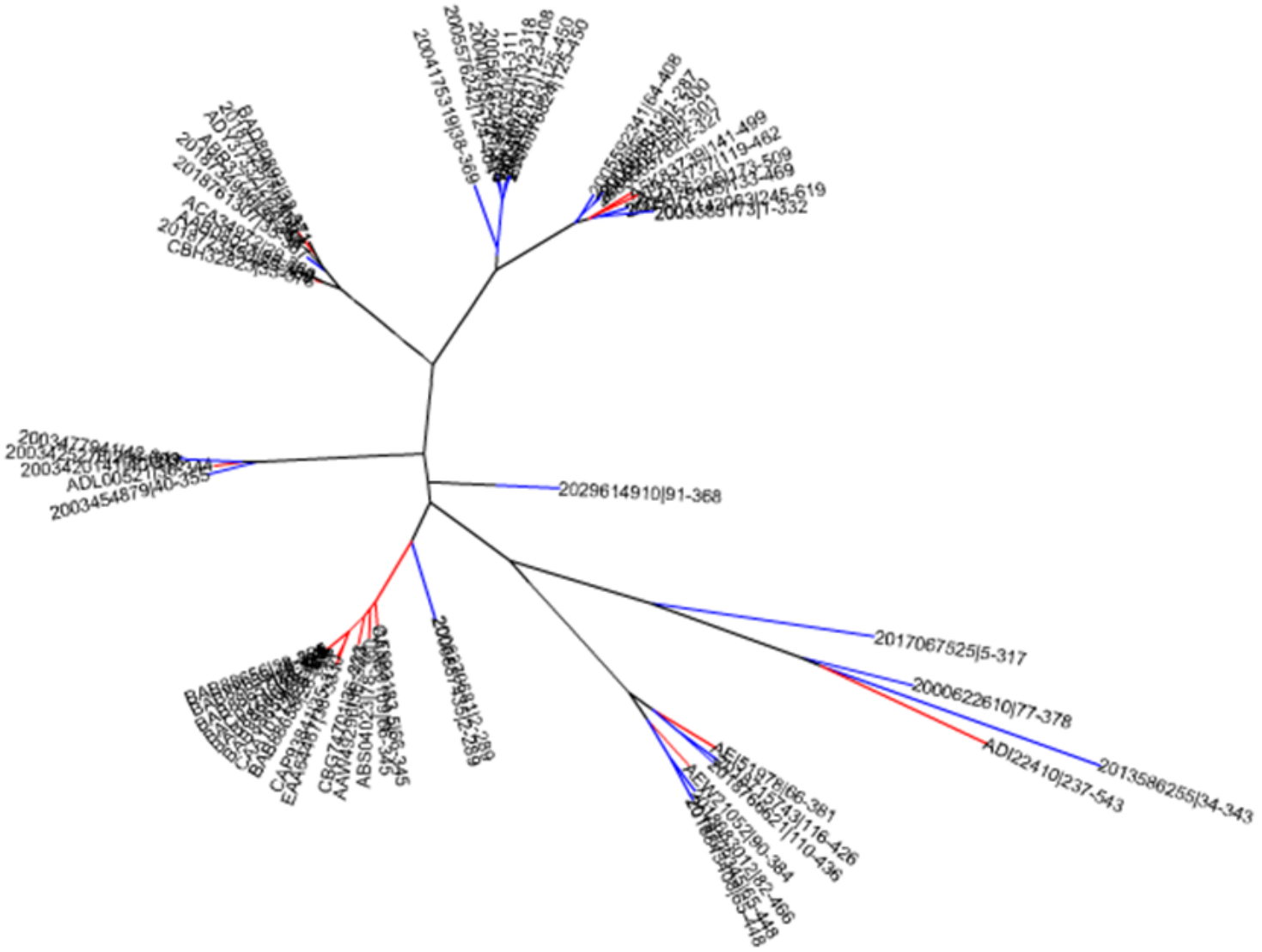
MAJOR FINDINGS


Datasets
Color

Pruning Phylogenetic Tree



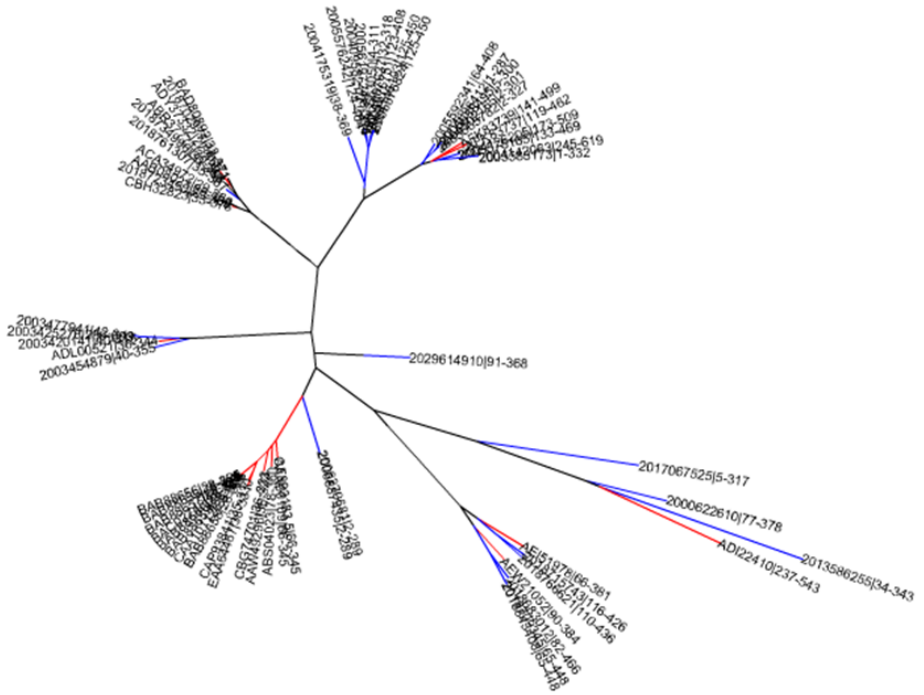
Pruned tree with JGI metagenome proteins



What do we take away from this?

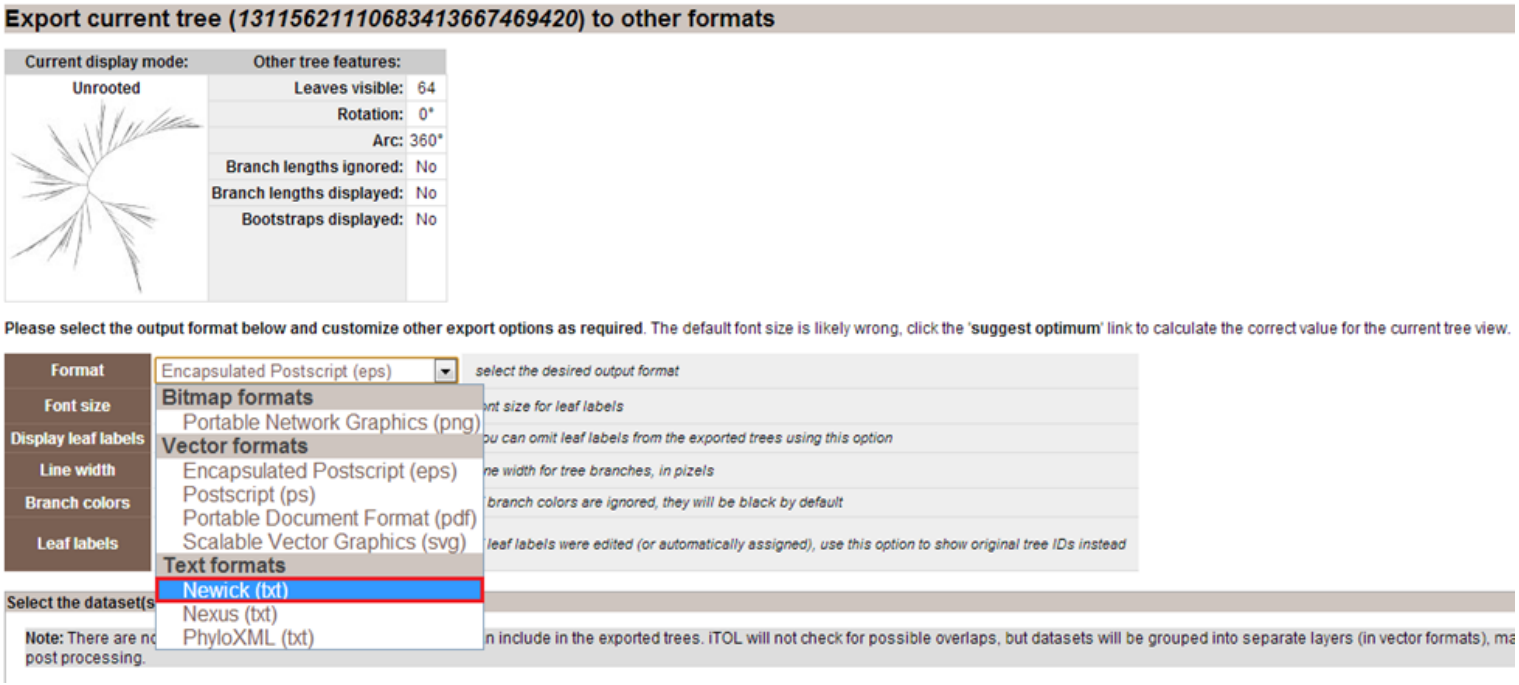
RESULTS!

- A few metagenome proteins share stringent sequence homology with the HMM domains of known GH10 proteins



- What are the JGI metagenome proteins that share stringent sequence homology with known GH10 proteins?

Export pruned tree form iTOL



Exporting the tree into newick format

Your file is ready for download. It will be kept on our server for 1 week.

Citations: Letunic and Bork (2006) *Bioinformatics* **23(1):127-8** and Letunic and Bork (2011) *Nucleic Acids Res* **doi: 10.1093/nar/gkr201**

((((2029614910|91-368:0.43034):0.07385[0.87]):0.16780[0.99]):0.12117[0.95],((((((((((AEW21052|90-384:0.41615,(2018683012|82-466:0.28185,(2018709345|65-448:0.0,2018843408|65-448:0.0):0.32495):0.28572[1.00]):0.21689[0.97],(AEI51978|66-381:0.48947,(2018715743|116-426:0.33270,2018766621|110-436:0.32460):0.21405[0.98]):0.19183[0.95]):0.47197[1.00]):0.08954[0.84]):0.42771[0.99]):0.19136[0.72],((2017067525|5-317:1.51291,(2000622610|77-378:0.77634,(2013586255|34-343:1.96925,ADI22410|237-543:1.22077):0.15066[0.06]):1.06110[0.98]):0.71364[0.94]):0.25268[0.85]):0.33104[0.93]):0.21310[0.93]):0.11235[0.07],((2006270681|2-289:0.0,2006887435|2-289:0.0):0.53870,((((AAW49296|36-333:0.0,CBG74701|36-333:0.0):0.19803,((((((BAB88655|28-325:0.0,BAB88656|28-325:0.0):0.00311,BAB88657|28-325:0.00622):0.00309[0.77],(BAB88660|28-325:0.0,BAC02742|28-325:0.0):0.00939):0.00309[0.78],(CAA10112|28-325:0.00014,BAB88659|28-325:0.00623):0.00310[0.80]):0.00310[0.76],BAB88658|28-325:0.00014):0.18357[1.00],(EAA64467|38-331:0.14702,CAP93841|35-330:0.09824):0.08608[0.96]):0.14322[0.99]):0.06268[0.85],ABS04023|78-360:0.19844):0.05410[0.38],(AF393183_5|66-345:0.0,CAN00109|66-345:0.0):0.16617):0.47162[1.00]):0.28792[0.99]):0.02786[0.19]):0.10541[0.91],((((((((((((2004175319|38-369:0.44529,((2005563050|4-311:0.06525,((2004087637|32-318:0.00910,2005576751|123-408:0.00321):0.03897[0.99],(2004146193|125-450:0.00962,2005576824|125-450:0.01532):0.02946[0.95]):0.06334[1.00]):0.06137[0.96],(2005576242|124-404:0.14891,(2004095821|21-315:0.00014,2005619724|6-294:0.00014):0.08218[1.00]):0.04073[0.78]):0.31556[1.00]):0.04932[0.26]):0.09091[0.95],((((((2004126205|173-509:0.0,2005578165|133-469:0.0):0.20382,(2004142063|245-619:0.00014,2005583173|1-332:0.00014):0.37234[1.00]):0.06424[0.79],(AFK83739|141-499:0.30058,AFK83737|119-462:0.33129):0.02341[0.06]):0.09593[0.99],(2005592341|64-408:0.17191,((2004096412|1-287:0.00853,2005588495|5-300:0.01095):0.09764[1.00],(2004102165|2-301:0.02596,2005565782|2-327:0.01088):0.04402[0.95]):0.11588[1.00]):0.05234[0.79]):0.03978[0.83]):0.40091[1.00]):0.16616[0.91]):0.03618[0.39]):0.13718[0.94]):0.18219[0.98]):0.08241[0.52]):0.31[0.79]):0.13355[0.95],((((((((((CBH32823|33-373:0.06338,(2018723490|29-369:0.00014,(AAB08023|29-369:0.0,ACA34972|29-369:0.0):0.00014):0.08299[1.00]):0.11338[1.00](2018761307|35-367:0.16078,((ADY37332|27-365:0.20177,(2018702977|32-371:0.00491,BAD80892|32-371:0.01077):0.11089[1.00]):0.07301[0.96],(2018734996|30-368:0.00014,ABR37771|30-368:0.00014):0.19162[1.00]):0.04576[0.39]):0.14668[1.00]):0.08367[0.84]):0.03847[0.80]):0.06794[0.41]):0.09777[0.97]):0.10536[0.96]):0.03786[0.80]):0.03562[0.78]):0.46[0.93]):0.15611[0.97]):0.09306[0.95]):0.05279[0.86]):0.04635[0.25]):0.03361[0.57]):0.08321[0.93]):0.10021[0.68]):0.27754[1.00],(((((2003454879|40-355:0.33684,(ADL00521|36-344:0.19011,(2003420141|40-315:0.02196,(2003425276|296-603:0.02465,2003477941|42-349:0.01723):0.03078[0.89]):0.36598[1.00]):0.09747[0.74]):0.12927[0.93]):0.60420[1.00]):0.13685[0.89]):0.10903[0.70]):0.12973[0.83]):0.19026[0.99]):0.08397[0.82]);

```
z117479@glu:~/project1$ vi GH10.hmm.jgi.NCBI-nr.itolprune.nwk
050|4-311:0.06525,((2004087637|32-318:0.00910,2005576751|123-408:0.00321):0.0389
7[0.99],(2004146193|125-450:0.00962,2005576824|125-450:0.01532):0.02946[0.95]):0
.06334[1.00]):0.06137[0.96],(2005576242|124-404:0.14891,(2004095821|21-315:0.000
14,2005619724|6-294:0.00014):0.08218[1.00]):0.04073[0.78]):0.31556[1.00]):0.0493
2[0.26]):0.09091[0.95],((((((2004126205|173-509:0.0,2005578165|133-469:0.0):0.20
382,(2004142063|245-619:0.00014,2005583173|1-332:0.00014):0.37234[1.00]):0.06424
[0.79],(AFK83739|141-499:0.30058,AFK83737|119-462:0.33129):0.02341[0.06]):0.0959
3[0.99],(2005592341|64-408:0.17191,((2004096412|1-287:0.00853,2005588495|5-300:0
.01095):0.09764[1.00],(2004102165|2-301:0.02596,2005565782|2-327:0.01088):0.0440
2[0.95]):0.11588[1.00]):0.05234[0.79]):0.03978[0.83]):0.40091[1.00]):0.16616[0.9
1]):0.03618[0.39]):0.13718[0.94]):0.18219[0.98]):0.08241[0.52]):0.18631[0.79]):0
```

Copy and paste newick file to terminal using vi

Extract tree ID's using bioperl modules

```
z117479@glu:~/project1$ vi get.treeid.pl
```

```
#!/usr/bin/perl -w

use Bio::TreeIO;

$treeio=Bio::TreeIO->new(-format=>"newick", -file=>$ARGV[0]);

while($tree=$treeio->next_tree){
        for $node ($tree->get_nodes){
        print $node->id,"\n";
        }
}
```

Run perl script on newly created newick file from pruned iTOL tree

```
z117479@glu:~/project1$ perl get.treeid.pl GH10.hmm.jgi.NCBI-nr.itolprune.nwk |
sed '/^$/d' > GH10.hmm.jgi.NCBI-nr.itolprune.id
z117479@glu:~/project1$ less GH10.hmm.jgi.NCBI-nr.itolprune.id
```

```
2029614910|91-368
AEW21052|90-384
2018683012|82-466
2018709345|65-448
2018843408|65-448
AEI51978|66-381
2018715743|116-426
2018766621|110-436
2017067525|5-317
2000622610|77-378
2013586255|34-343
ADI22410|237-543
2006270681|2-289
2006887435|2-289
AAW49296|36-333
CBG74701|36-333
BAB88655|28-325
BAB88656|28-325
BAB88657|28-325
BAB88660|28-325
BAC02742|28-325
CAA10112|28-325
BAB88659|28-325
GH10.hmm.jgi.NCBI-nr.itolprune.id
```

Parse output file to remove sequence coordinates

```
z117479@glu:~/project1$ less GH10.hmm.jgi.NCBI-nr.itolprune.id | sed 's/|/\t/' |
 cut -f1 > GH10.hmm.jgi.NCBI-nr.itolprune.id.ps
z117479@glu:~/project1$ less GH10.hmm.jgi.NCBI-nr.itolprune.id.ps
```

```
2013586255
ADI22410
2006270681
2006887435
AAW49296
CBG74701
BAB88655
BAB88656
BAB88657
BAB88660
BAC02742
CAA10112
BAB88659
GH10.hmm.jgi.NCBI-nr.itolprune.id.ps
```

Works Cited

Béguin, Pierre. "MOLECULAR BIOLOGY OF CELLULOSE DEGRADATION." *Annu. Rev. Microbiol* 4 (1990): 219-48. Print.

HARRIS, GILLIAN W. "Refined Crystal Structure of the Catalytic Domain of Xylanase A from Pseudomonas Fluorescens at 1.8 .~ Resolution." *Acta Crystallographica Section D/ International Union of Crystallography* D52 (1996): 393-401. Web. < http://journals.iucr.org/d/issues/1996/02/00/he0136/he0136.pdf >.

"InterPro." *Glycoside Hydrolase, Family 10*. EMBL-EBI, n.d. Web. <http://www.ebi.ac.uk/interpro/entry/IPR001000>.

Manow, Ryan. "Partial Deletion of Rng (RNase G)-enhanced Homoethanol Fermentation of Xylose by the Non-transgenic Escherichia Coli RM10." *J Ind Microbiol Biotechnol* 39 (2012): 977-85. Print.

Withers, Steven. "Glycoside Hydrolase Family 10." *CAZypedia*. N.p., 10 Sept. 2012. Web. <http://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_10#bibkey_HJG>.

Withers, Steve, and Spencer Williams. "Glycoside Hydrolases." *CAZypedia*. N.p., 15 Jan. 2013. Web. <http://www.cazypedia.org/index.php/Retaining>.

"Xylanase." *Wikipedia*. Wikimedia Foundation, 19 Apr. 2013. Web. <http://en.wikipedia.org/wiki/Xylanase>.