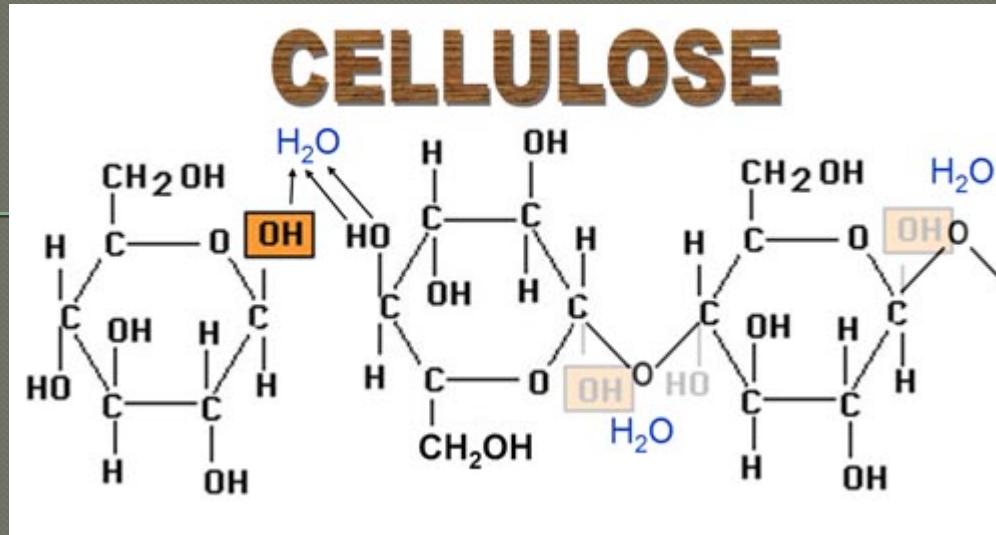


A Search for Novel GH9 Homologues within a Metagenomic Database

Tom Bean
Brenda Pierson
Steve Seydell
Bill Wysocki

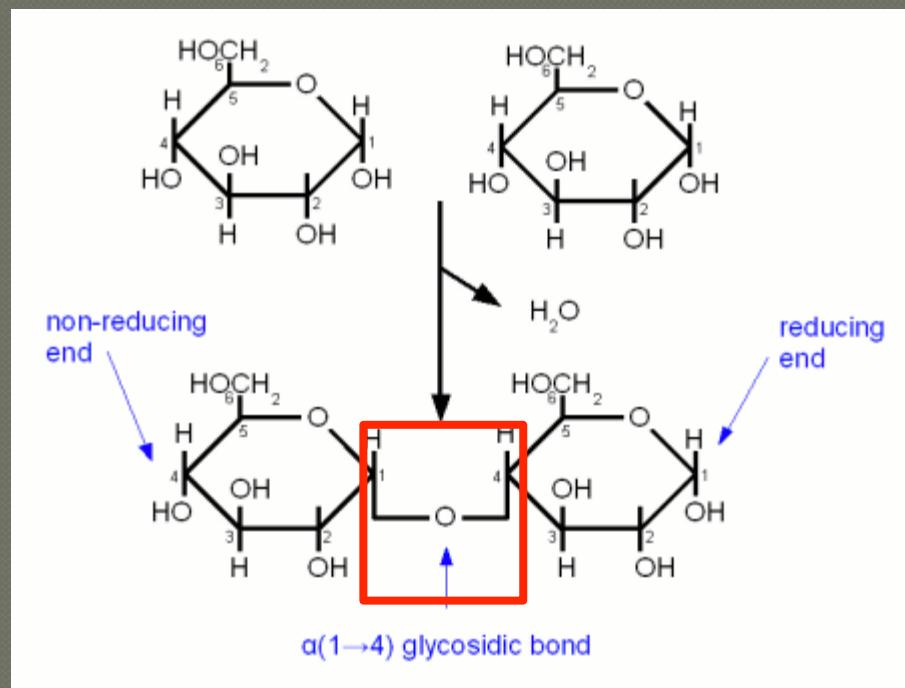


- Polysaccharide found in plant cell walls.
- Very numerous in nature
- Broken down by cellulase
 - Defined by using cellulose as a substrate.

Sukharnikov et al., 2012

Glycoside Hydrolases

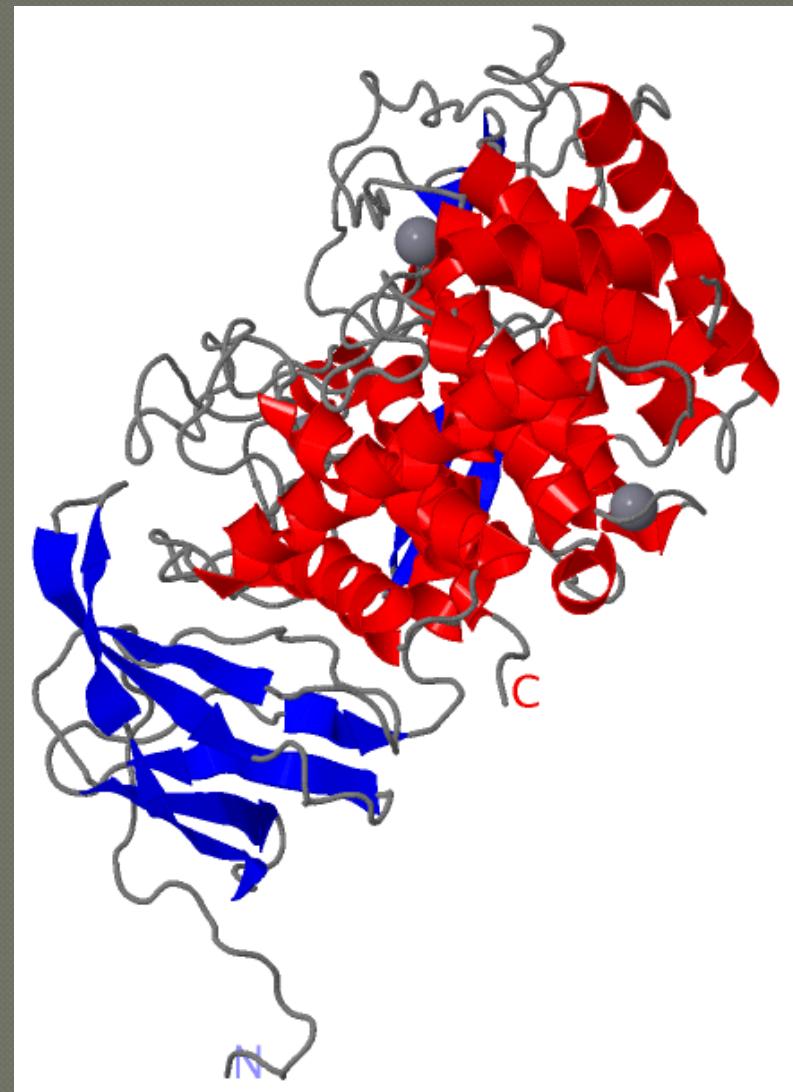
- Break glycosidic linkages
- Understanding is crucial to understanding the ‘CAZome’



Aspeborg et al., 2012

Glycoside hydrolase (GH) family 9

- All known plant cellulases belong to GH9
- 2nd largest cellulase family
- Most plant GH9 enzymes studied are cellulases



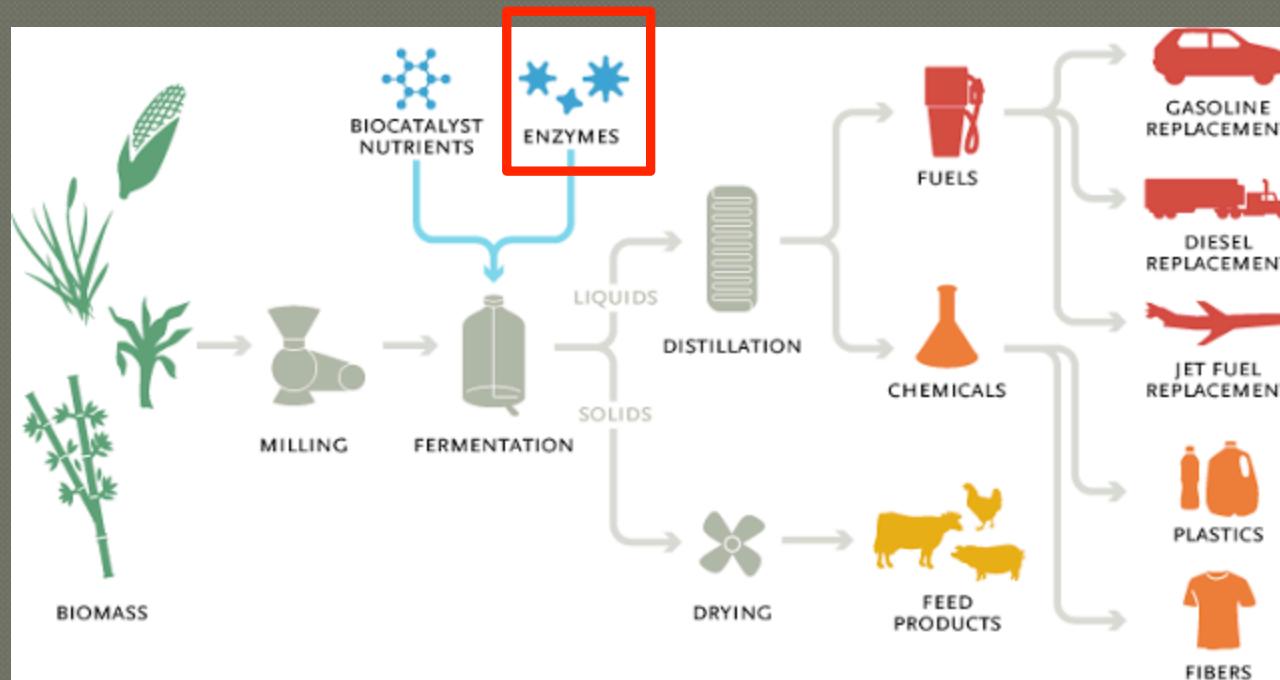
In Nature...

- Cellulases are involved in breakdown of dead plant cells
- Plants need cellulases for shedding of tissues, fruit expansion, organ development



Why do we care?

Renewable chemicals company **Gevo** is now producing a range of products from biomass cellulose-derived sugars.



Gevo first ferments isobutanol from biomass (cellulose-derived sugars), then converts the isobutanol to other high value materials.

Project Goal:

**Here we intend to identify novel
GH9 cellulase domain families
found in publicly-available
metagenomic sample databases**

PROCEDURE OUTLINE

CAZy

Accessions

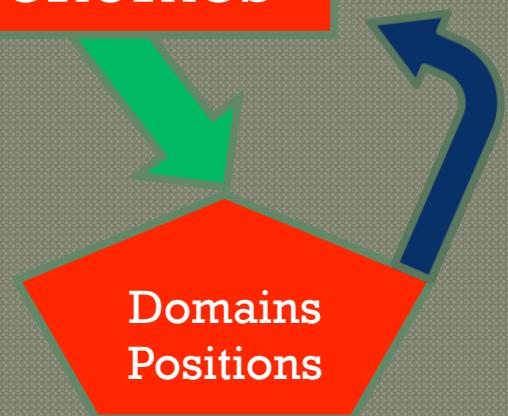
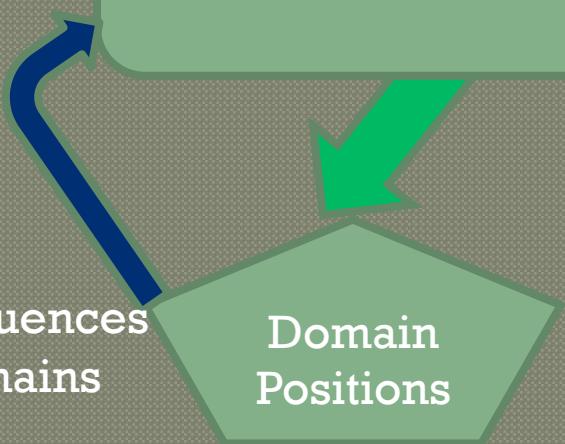
NCBI

GH9 HMM

HMMSEARCH

Known GH9 Protein FASTA
sequences from NCBI

JGI
 >1000
metagenomes



CAZy

Accessions

NCBI

GH9 HMM

HMMSEARCH

Known GH9 Protein FASTA
sequences from NCBI

JGI
 >1000
metagenomes

Domain
Positions

Domains
FASTA

Domains
Positions

- Filter
- Extract sequences
- Extract domains

Domains
FASTA

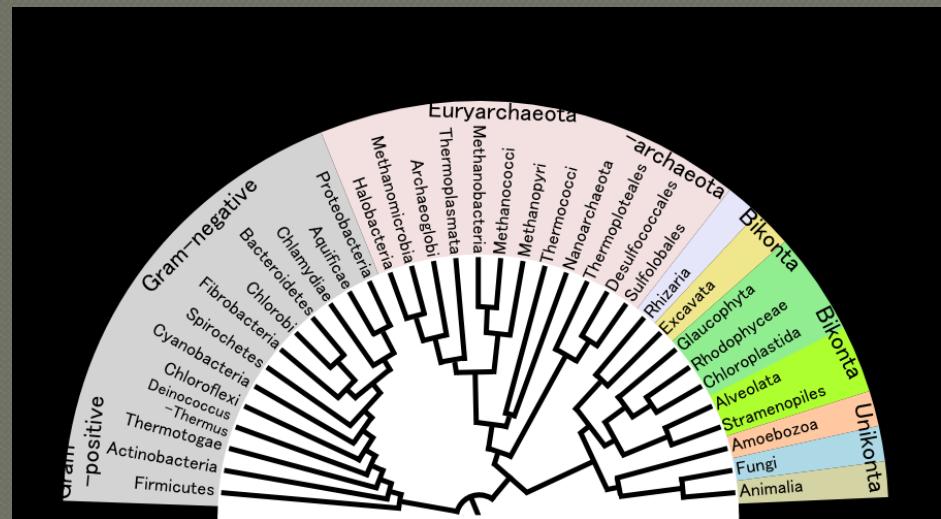
JGI: 1777 GH9 domains from 211 metagenomes
NCBI 1452 GH9 domains

MAFFT

Domains
Aligned

FastTree

Phylogeny



ITOL

CAZy

Accessions

NCBI

GH9 HMM

HMMSEARCH

Known GH9 Protein FASTA
sequences from NCBI

JGI
 >1000
metagenomes

Domain
Positions

Domains
FASTA

Domains
Positions

- Filter
- Extract sequences
- Extract domains

Dataset Download

Summary	All (1518)	Archaea (2)	Bacteria (530)	Eukaryota (779)	unclassified (207)	Structure (11)	Characterized (132)
< 1 2 >							
Archaea							
Protein Name							
Hmuk_0843							GenBank
Huta_2399							ACV46974.1
Bacteria							
Protein Name							
AM1_2816							GenBank
scaffoldin A (ScaA;CipV)							ABW27815.1
endo-β-1,4-glucanase (Cel9B)							AAF06064.1
ACP_1014							CAI94607.1
Acel_0970							ACG22406.1
Acel_1701							ABK52743.1
AMIS_26270							ABK53473.1
AMIS_35140							BAL87847.1
AMIS_48350							BAL88734.1
ACPL_1206							BAL90055.1
ACPL_2890 (Cel1)							AEV82103.1
Amir_5339							AEV83785.1
AHA_3098							ACU39159.1
β-glucosidase							ABK39743.1
ASA_3105							BAF75999.1
B565_1044							ABO91099.1
cellulase (CelA;Aaci_2475) (Cel9A)							AEB49079.1
...							ACV59481.1
							CAC34051.1

•MS-Excel

•less GH9.accessions.raw | sed '/^\$/d' | grep [0-9] > GH9.accessions.clean

•Genbank Query

Extraction from NCBI

Batch Entrez

2092 Known Proteins

```
>gi|20415|emb|CAA39314.1| cellulase [Persea americana]
ASCGSTTVAKNLISLAKKQVDYILGENPAKMSYMVGFGERYPQHVHHRGSSLPSVAHPNPIPCNAGFQ
YLYSSSPNPNILVGAILGGPDSRDSFSDDRNNYQQSEPATYINAPLVGALAFFAANPVAN

>gi|20417|emb|CAA42569.1| cellulase [Persea americana]
MDCSSPLSLFHLLVCTVMVKCCSASDLHYSDALEKSILFFEGQRSGKLPTNQRLTWRGDGLSDGSSYH
VDLVGGYYDAGDNLKFGLPMAFTTMLAWGIIEFGCLMPEQVENARAALRWSTDYLLKASTATSNSLYVQ
VGEPNADHRCWERPEDMDTPRNVYKVSTQNPGSDVAAETAALAAASIVFGDSDSSYSTKLLHTAVKFE
FADQYRGYSSDLGSVVCPFYCSGYNDELLWGASWLHRASQNASYMTYIQSNGHTLGADDDYSFSWD
DKRVGTVKVLLSKGFLQDRIEELQLYKVHTDNYICSLIPGTSSFQAQYTPGGLLYKGSASNQYVTSTAFL
LLTYANYLNSSGGHASCFTTVAKNLISLAKKQVDYILGQNPAKMSYMVGFGERYPQHVHHRGSSLPSV
QVHPNSIPCNAQFQYLYSSPPNPNILVGAILGGPDNRDSFSDDRNNYQQSEPATYINAPLVGALAFFAAN
PVTE

>gi|40672|emb|CAA28255.1| unnamed protein product [Clostridium thermocellum]
MSRMTLKSSMKKRVLSSLIAVVFLSLTGVFPSGLIETKVSAAKITEINYQFDUSRIRLN SIGFIPNHSKKAT
IAANCSTFYVVKEDGTIVYTGTATSMFDNDTKETVYIADFSSVNEEGTYYLAVPGVGKSVNFKIAMNVYE
DAFKTAMLGMYLLRCGTSVSATYNGIHSHGPCHTN DAYLDYINGQHTKKDSTKGWH DADGDYNKYVVNAG
ITVGSMFLAWEHKDQLEPV ALEIPEKNNSIPDFLDELKYEIDWILTMQYPDGSGRVAHKVSTRNFGGF
MPENEHDERFFVFWSSAATADFVAMTAMAARIFRPYDPQYAECKCINA AKVSYEFLKNNPANVFA NQSGFS
TGEYATVSDADDRLWAAAEMWETLGDEEYL RDFENRAAQFSKKIEADFDWDNVANLGMFTYLLSERPGKN
PALVQSIKDSL LSTADSIVRTSQNHGYGRTL GTTYYWGCNGTVVRQTMILQVANKISPNDYVNA ALDAI
SHVFGRNYYNRSYVTGLGINPPMNPHDRRSGADGIWE PWPGYLVGGGWPGPKDWVDI QDSYQTNEIA INW
gh9.fasta
```

CAZy

Accessions

NCBI

GH9 HMM

HMMSEARCH

Known GH9 Protein FASTA
sequences from NCBI

2092

JGI
>1000
metagenomes

- Filter
- Extract sequences
- Extract domains

Domain
Positions

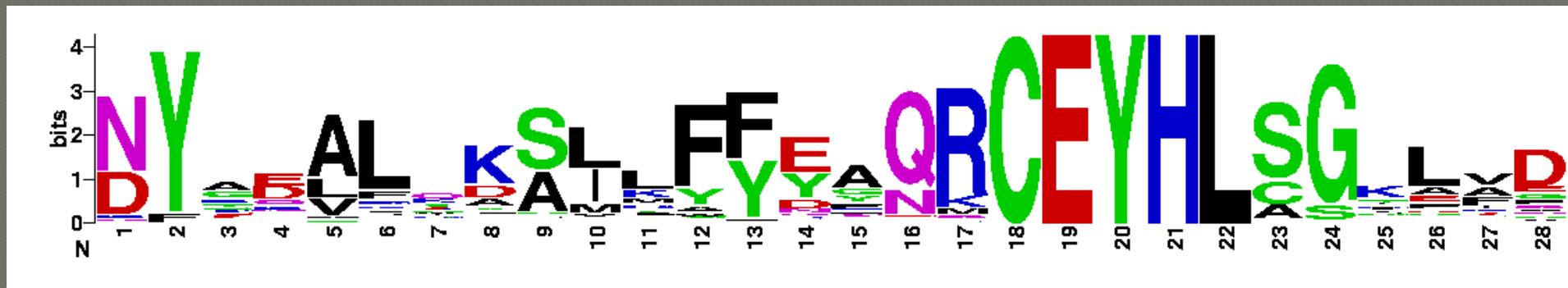
Domains
FASTA

Domains
Positions

Command lines used to do homology search

```
hmmsearch --domtblout GH9.NCBI.HMM.out GH9.hmm.ps NCBIbatchentrez.fasta &
```

```
hmmsearch --domtblout GH9.JGI.HMM.out GH9.hmm.ps /home/yyin/db/jgi-v3.5.pr.fa &
```

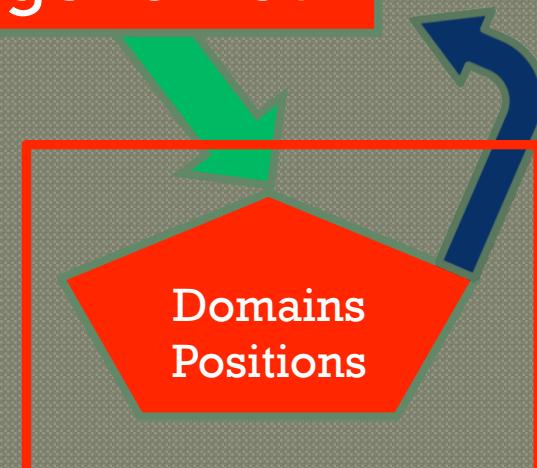
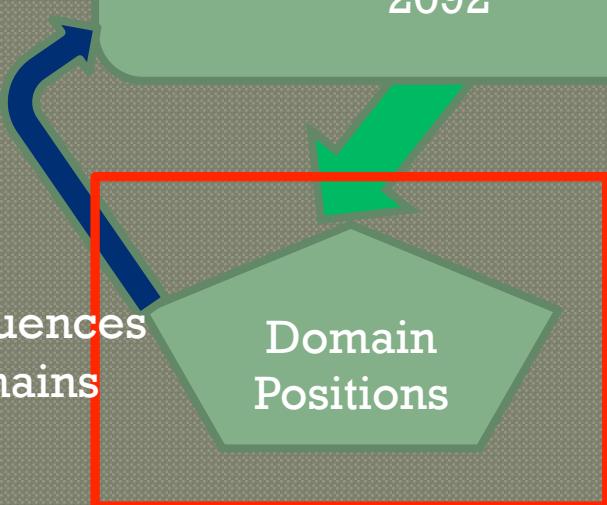


CAZy
Accessions



Known GH9 Protein FASTA
sequences from NCBI
2092

JGI
>1000
metagenomes



- Filter
- Extract sequences
- Extract domains

GH9 NCBI hmm - 2,102 Hits

domain -----	hmm	coord	ali	coord	env	coord	--- full sequence ---									
target name				accession	tlen	query name	accession	qlen	E-value	score	bias	# of	c-Eva			
value	score	bias	from	to	from	to	acc	description of target								
i 42601450 gb AAS21473.1	.9e-107	356.2	6.5	1	417	53	479	1000 GH9.hmm	-	418	4.7e-193	639.2	21.3	1	2	3.9e
i 42601450 gb AAS21473.1	.2e-87	289.2	2.8	1	417	533	963	533 964 0.90 beta-1,4-endoglucanase 1 [Oikopleura dioica]	-	418	4.7e-193	639.2	21.3	2	2	8.2e
i 371940140 dbj BAL45507.1	.7e-152	503.9	18.0	1	418	41	475	41 475 0.94 glycoside hydrolase [Bacillus licheniformis]	-	418	4.7e-152	504.1	26.0	1	1	5.7e
i 264670906 gb ACY72380.1	.6e-150	499.1	14.6	1	418	34	467	34 467 0.93 cellulose hydrolase [Bacillus licheniformis]	-	418	1.1e-150	499.6	21.1	1	1	1.6e
i 39636954 gb AAR29083.1	.1e-150	498.2	14.9	1	418	21	454	21 454 0.93 cellulase [Bacillus licheniformis]	-	418	2.4e-150	498.5	21.5	1	1	3.1e
i 52003473 gb AAU23415.1	.4e-150	498.0	14.9	1	418	42	475	42 475 0.93 Glycoside Hydrolase Family 9 [Bacillus licheniformis DSM 13 = ATCC	-	418	2.7e-150	498.3	21.5	1	1	3.4e
i 52348141 gb AAU40775.1	.4e-150	498.0	14.9	1	418	42	475	654 GH9.hmm	-	418	2.7e-150	498.3	21.5	1	1	3.4e
i 4490766 emb CAB38941.1	.9e-149	494.9	18.8	1	418	40	479	42 475 0.93 cellulase EglA [Bacillus licheniformis DSM 13 = ATCC 14580]	-	418	1.5e-149	495.9	27.1	1	1	2.9e
i 37498962 gb AAQ91573.1	.7e-149	494.2	13.8	1	418	48	480	997 GH9.hmm	-	418	3.7e-149	494.6	19.9	1	1	4.7e
i 374343072 dbj BAL46914.1	.2e-149	493.3	17.5	1	418	40	479	48 480 0.94 endoglucanase A precursor [Bacillus pumilus]	-	418	4.8e-149	494.2	25.2	1	1	9.2e
i 188011009 gb ACD44896.1	.9e-148	492.2	13.7	1	418	5	437	616 GH9.hmm	-	418	1.6e-148	492.5	19.8	1	1	1.9e
i 310751795 gb ADP09350.1	.9e-148	492.2	13.3	1	418	48	480	5 437 0.94 endoglucanase [Bacillus pumilus]	-	418	1.6e-148	492.5	19.2	1	1	1.9e
i 148767913 gb ABR10904.1	.4e-148	491.9	13.7	1	418	48	480	659 GH9.hmm	-	418	1.9e-148	492.2	19.8	1	1	2.4e
i 6525242 gb AAF15367.1	.4e-148	491.9	13.7	1	418	48	480	48 480 0.94 endo-1,4-beta-glucanase [Bacillus pumilus]	-	418	1.9e-148	492.2	19.8	1	1	2.4e
i 122937813 gb ABM68635.1	.4e-148	491.9	13.7	1	418	48	480	659 GH9.hmm	-	418	1.9e-148	492.2	19.8	1	1	2.4e
i 1817723 gb AAB42155.1	.2e-148	490.6	13.6	1	418	51	484	48 480 0.94 endoglucanase A [Bacillus sp. AC-1]	-	418	4.9e-148	490.9	19.6	1	1	6.2e
i 71916307 gb AAZ56209.1	.2e-148	490.6	13.6	1	418	51	484	880 GH9.hmm	-	418	4.9e-148	490.9	19.6	1	1	6.2e
i 42601450 gb AAS21473.1	.2e-148	490.6	13.6	1	418	51	484	880 GH9.hmm	-	418	4.9e-148	490.9	19.6	1	1	6.2e
i 42601450 gb AAS21473.1	.2e-148	490.6	13.6	1	418	51	484	51 484 0.94 endoglucanase. Glycosyl Hydrolase family 9 [Thermobifida fusca YX]	-	418	4.9e-148	490.9	19.6	1	1	6.2e

GH9 JGI hmm - 15,428 Hits

#	env coord	target name	accession	tlen	query name	accession	qlen	--- full sequence ---			this domain			hmm	coord	ali	coord		
	from	to	acc	description of target				E-value	score	bias	# of c-Value	i-E-value	score	bias	from	to	from	to	
2201231693	-	883	GH9.hmm	-		418	2.9e-146	501.1	17.2	1	1	2.9e-150	3.6e-146	500.8	12.0	1	418	278	706
278	706	0.94	MRSJC2b_43491	Predicted solute binding protein [Mesophilic rice straw/compost enrichment metagenome: eDNA_1 (Mesophilic 454/Illumina Combined June 2011 assem)]															
2200984066	-	756	GH9.hmm	-		418	3.9e-144	494.1	24.0	1	1	4.1e-148	5.1e-144	493.7	16.6	1	418	37	474
37	474	0.93	MRSJC2b_272891	Glycosyl hydrolase family 9./Cellulose binding domain. [Mesophilic rice straw/compost enrichment metagenome: eDNA_1 (Mesophilic 454/Illumina Combined June 2011 assem)]															
2061998357	-	768	GH9.hmm	-		418	2.1e-143	491.7	24.1	1	1	2e-147	2.6e-143	491.4	16.7	1	418	51	488
51	488	0.92	sg4i_00246050	Predicted solute binding protein [Thermophilic enrichment culture SG0.5JP960 (454-Illumina assembly) - version 2 (454-Illumina assembly v2)]															
2018727539	-	880	GH9.hmm	-		418	3.6e-143	490.9	19.6	1	1	3.6e-147	4.5e-143	490.6	13.6	1	418	51	484
51	484	0.94	PFMN_291642	hypothetical protein [Sample 267]															
2201115235	-	1283	GH9.hmm	-		418	6.8e-141	483.4	15.8	1	1	7.6e-145	9.5e-141	482.9	11.0	1	418	666	1094
666	1094	0.94	MRSJC2b_32941	Uncharacterized protein conserved in bacteria [Mesophilic rice straw/compost enrichment metagenome: eDNA_1 (Mesophilic 454/Illumina Combined June 2011 assem)]															
2200446424	-	1001	GH9.hmm	-		418	4.9e-140	480.6	17.9	1	1	6.3e-144	7.9e-140	479.9	12.4	1	417	95	528
95	529	0.94	TRSJC2b_5223	Cellulobiohydrolase A (1,4-beta-celllobiosidase A) [Thermophilic rice straw/compost enrichment metagenome: eDNA_2 (Thermophilic 454/Illumina Combined June 2011 assem)]															
2078025289	-	1028	GH9.hmm	-		418	4.6e-139	477.4	22.8	1	1	5.9e-143	7.4e-139	476.7	15.8	1	418	73	503
73	503	0.93	MA55A_00355410	Cellulose binding domain./Glycosyl hydrolase family 9./Domain of unknown function. [Mixed alcohol bioreactor microbial communities from Texas A&M University, sample 55C (55 degree reactor, August 2010 assembly)]															
2029678661	-	569	GH9.hmm	-		418	1.5e-138	475.7	21.2	1	1	1.4e-142	1.8e-138	475.4	14.7	1	417	41	474
41	475	0.95	GBAN_2580090	Cellulose binding domain./Glycosyl hydrolase family 9. [Compost Minireactor Metagenome (final assembly)]															
2201670307	-	668	GH9.hmm	-		418	3.1e-138	474.6	21.3	1	1	3e-142	3.7e-138	474.4	14.8	1	418	30	469
30	469	0.94	MRSJC2b_190371	Glycosyl hydrolase family 9./Cellulose binding domain. [Mesophilic rice straw/compost enrichment metagenome: eDNA_1 (Mesophilic 454/Illumina Combined June 2011 assem)]															
2150562325	-	513	GH9.hmm	-		418	1.9e-137	472.0	28.9	1	1	1.8e-141	2.2e-137	471.8	20.1	1	418	36	466
36	466	0.91	GBSCES77_01292830	Glycosyl hydrolase family 9. [Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment S 77C (Cellulolytic enrichment S 77C sediment)]															
2150657657	-	513	GH9.hmm	-		418	1.9e-137	472.0	28.9	1	1	1.8e-141	2.2e-137	471.8	20.1	1	418	36	466
36	466	0.91	GBSCES77_02246150	Glycosyl hydrolase family 9. [Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment S 77C (Cellulolytic enrichment S 77C sediment)]															
2150468268	-	640	GH9.hmm	-		418	3.3e-137	471.3	28.9	1	1	3.6e-141	4.6e-137	470.8	20.1	1	418	36	466
36	466	0.91	GBSCES77_00352260	Cellulose binding domain./Glycosyl hydrolase family 9. [Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment S 77C (Cellulolytic enrichment S 77C sediment)]															
2150635057	-	640	GH9.hmm	-		418	3.3e-137	471.3	28.9	1	1	3.6e-141	4.6e-137	470.8	20.1	1	418	36	466
36	466	0.91	GBSCES77_02020150	Cellulose binding domain./Glycosyl hydrolase family 9. [Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment S 77C (Cellulolytic enrichment S 77C sediment)]															
2150588393	-	641	GH9.hmm	-		418	3.3e-137	471.3	28.9	1	1	3.7e-141	4.6e-137	470.8	20.1	1	418	36	466
36	466	0.91	GBSCES77_01553510	Cellulose binding domain./Glycosyl hydrolase family 9. [Sediment microbial communities from Great Boiling Spring, Nevada, sample from															

PlainText ▾ Tab Width: 8 ▾ Ln 1, Col 1 IN:

CAZy

Accessions

NCBI

GH9 HMM

HMMSEARCH

Known GH9 Protein FASTA
sequences from NCBI

2092

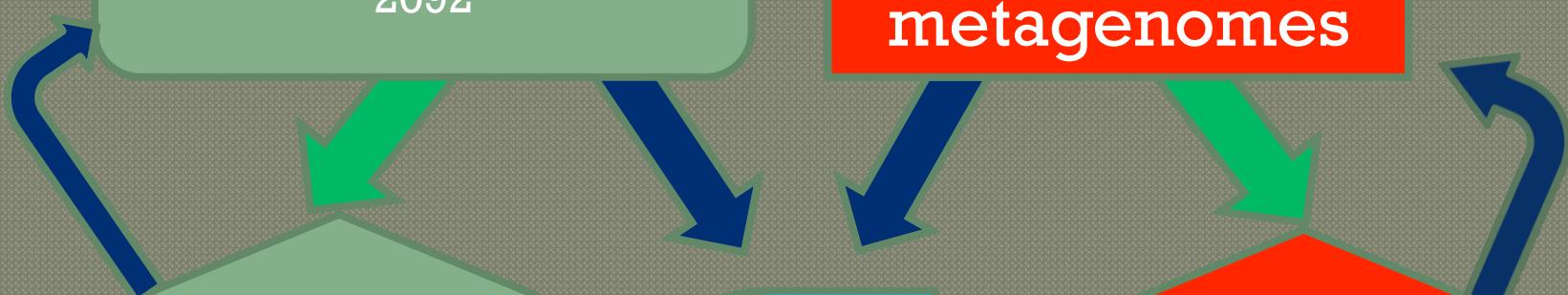
JGI
>1000
metagenomes

- Filter
- Extract sequences
- Extract domains

Domain
Positions

Domains
FASTA

Domains
Positions



Perl script to:
extract hit id,e-value, bit score,query start, query end, hit start,
hit end as a tabular file.
Filter by E-values <1X10⁻⁵, Coverage over 80%

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

#extract the hit id, e-value, bit score, query start, query end, hit
#start, and hit end as a tabular file

open(OUT, ">GH9.ncbi.hmm.out.4.1.1");
open(IN,$ARGV[0]);
print OUT "Hit ID\t\t\te-value\t\tbit score\tquery start\tquery end\thit start\thit end\n";
while(<IN>){
    if($_ =~/^#/){
        $_=~s/^$/; #first if statement removes the header lines of the .out #file
    }
    else{
        @line = $_; #stores all the lines into an array
    }

    foreach(@line){
        #for each line there is an array created for columns which are #separated by whitespaces
        @col=split(/\s+/,$_);

        #check to see if e-value is less than 1e-5 and also has a coverage >0.8
        if($col[6]<1e-5&&((($col[16]-$col[15])/$col[5])>0.8){
            #if true, prints out the required values.
            print OUT $col[0], "\t", $col[6], "\t", $col[7], "\t\t\t", $col[15], "\t", $col[16], "\t", $col[17], "\t", $col[18], "\n";
        }
    }
}
```

GH9 JGI hmm out tabluar format

hit ID	e-value	bit score	query start	query end	hit start	hit end
201231693	2.9e-146	501.1	1	418	278	706
200984066	3.9e-144	494.1	1	418	37	474
061998357	2.1e-143	491.7	1	418	51	488
018727539	3.6e-143	490.9	1	418	51	484
201115235	6.8e-141	483.4	1	418	666	1094
200446424	4.9e-140	480.6	1	417	95	528
078025289	4.6e-139	477.4	1	418	73	503
029678661	1.5e-138	475.7	1	417	41	474
201670307	3.1e-138	474.6	1	418	30	469
150562325	1.9e-137	472.0	1	418	36	466
150657657	1.9e-137	472.0	1	418	36	466
150468268	3.3e-137	471.3	1	418	36	466
150635057	3.3e-137	471.3	1	418	36	466
150588393	3.3e-137	471.3	1	418	36	466
150644411	3.3e-137	471.3	1	418	36	466
150477690	3.3e-137	471.3	1	418	36	466
150525080	3.3e-137	471.3	1	418	36	466
150601450	3.3e-137	471.3	1	418	36	466
150611438	3.3e-137	471.3	1	418	36	466
150498305	3.3e-137	471.3	1	418	36	466
150563281	3.3e-137	471.3	1	418	36	466
077991057	2.8e-135	464.9	1	418	53	483
150675743	3e-134	461.5	1	418	42	460
150675975	3e-134	461.5	1	418	42	460
150635401	3e-134	461.5	1	418	42	460
150652215	3e-134	461.5	1	418	42	460
150447302	3.1e-134	461.5	1	418	42	460
150533972	3.1e-134	461.5	1	418	42	460
150558818	3.1e-134	461.5	1	418	42	460
150667124	3.1e-134	461.5	1	418	42	460
150471155	3.1e-134	461.5	1	418	42	460
150524468	3.1e-134	461.5	1	418	42	460
120604568	3.1e-134	461.5	1	418	40	469
018778051	2.4e-133	458.6	1	418	39	512
096696405	2.8e-133	458.3	1	418	97	515
120702712	1.3e-132	456.2	1	418	39	513
165286575	1.3e-132	456.2	1	418	39	513
018747411	4.1e-131	451.2	1	417	40	468
096800528	2.5e-130	448.6	1	418	117	537
018706652	4.3e-130	447.8	1	418	19	497
018741988	1.3e-126	436.4	1	418	41	511
018820060	1.3e-126	436.4	1	418	41	511
018770236	3.4e-126	435.0	1	418	30	451

GH9 JGI hit id's

1777

```
2201231693
2200984066
2061998357
2018727539
2201115235
2200446424
2078025289
2029678661
2201670307
2150562325
2150657657
2150468268
2150635057
2150588393
2150644411
2150477690
2150525080
2150601450
2150611438
2150498305
2150563281
2077991057
2150675743
2150675975
2150635401
2150652215
2150447302
2150533972
2150558818
2150667124
2150471155
2150524468
2120604568
2018778051
2096696405
2120702712
2165286575
2018747411
2096800528
2018706652
2018741988
2018820060
2018770236
2018691481
```

Plain Text ▾

GH9 NCBI hmm out tabular format

hit ID	e-value	bit score	query start	query end	hit start	hit end
i 42601450 gb AAS21473.1	4.7e-193	639.2	1	417	53	479
i 42601450 gb AAS21473.1	4.7e-193	639.2	1	417	533	963
i 371940140 dbj BAL45507.1	4.7e-152	504.1	1	418	41	475
i 264670906 gb ACY72380.1	1.1e-150	499.6	1	418	34	467
i 39636954 gb AAR29083.1	2.4e-150	498.5	1	418	21	454
i 52003473 gb AAU23415.1	2.7e-150	498.3	1	418	42	475
i 52348141 gb AAU40775.1	2.7e-150	498.3	1	418	42	475
i 4490766 emb CAB38941.1	1.5e-149	495.9	1	418	40	479
i 37498962 gb AAQ91573.1	3.7e-149	494.6	1	418	48	480
i 374343072 dbj BAL46914.1	4.8e-149	494.2	1	418	40	479
i 188011009 gb ACD44896.1	1.6e-148	492.5	1	418	5	437
i 310751795 gb ADP09350.1	1.6e-148	492.5	1	418	48	480
i 148767913 gb ABR10904.1	1.9e-148	492.2	1	418	48	480
i 6525242 gb AAF15367.1	1.9e-148	492.2	1	418	48	480
i 122937813 gb ABM68635.1	1.9e-148	492.2	1	418	48	480
i 1817723 gb AAB42155.1	4.9e-148	490.9	1	418	51	484
i 71916307 gb AAZ56209.1	4.9e-148	490.9	1	418	51	484
i 157681096 gb ABV62240.1	5.8e-148	490.7	1	418	5	437
i 2897802 dbj BAA24918.1	1.2e-147	489.6	1	415	29	458
i 144808 gb AAA20892.1	1.2e-147	489.6	1	418	77	509
i 125712789 gb ABN51281.1	1.2e-147	489.6	1	418	77	509
i 7208809 emb CAB76932.1	1.2e-147	489.6	1	418	77	509
i 316941198 gb ADU75232.1	2e-147	488.9	1	418	77	509
i 353441419 gb AEQ94264.1	2.5e-147	488.6	1	418	48	480
i 219544539 gb ACL26277.1	5.3e-147	487.5	1	418	31	467
i 451784663 gb AGF55631.1	5.3e-147	487.5	1	418	39	469
i 372100184 gb AEX68682.1	1.1e-146	486.4	1	418	5	437
i 141603871 gb ABO88214.1	1.3e-146	486.2	1	418	48	480
i 125713488 gb ABN51980.1	1.9e-146	485.7	1	418	31	463
i 316940505 gb ADU74539.1	1.9e-146	485.7	1	418	31	463
i 359824960 gb AEV67733.1	1.1e-145	483.2	1	418	30	460
i 99644508 emb CAK22316.1	1.5e-145	482.7	1	418	31	463
i 3600052 gb AAC35539.1	1.6e-145	482.7	1	418	26	487
i 4850284 emb CAB43040.1	1.6e-145	482.7	1	418	26	487
i 7267803 emb CAB81206.1	1.6e-145	482.7	1	418	26	487
i 22136600 gb AAM91619.1	1.6e-145	482.7	1	418	26	487
i 30681638 ref NP_192843.2	1.6e-145	482.7	1	418	26	487
i 332657566 gb AEE82966.1	1.6e-145	482.7	1	418	26	487
i 125713287 gb ABN51779.1	2.1e-145	482.3	1	418	31	461
i 316940717 gb ADU74751.1	2.1e-145	482.3	1	418	31	461
i 581006 emb CAA43035.1	2.1e-145	482.3	1	418	31	461
i 334812465 gb AEH04391.1	6.6e-145	480.6	1	418	51	484
i 6179388 emb CAB59900.1	3.8e-144	478.1	1	418	32	488

GH9 NCBI hit id's

1452

```
GH9.ncbi.nlm.nih.gov:1111/mcs... |x
L|42601|Close document|3.1|
L|42601456|gb|A52173.1|
L|371940140|dbj|BAL45507.1|
L|264670906|gb|ACY72380.1|
L|39636954|gb|AAR29083.1|
L|52003473|gb|AAU23415.1|
L|52348141|gb|AAU40775.1|
L|4490766|emb|CAB38941.1|
L|37498962|gb|AAQ91573.1|
L|374343072|dbj|BAL46914.1|
L|188011009|gb|ACD44896.1|
L|310751795|gb|ADP09350.1|
L|148767913|gb|ABR10904.1|
L|6525242|gb|AAF15367.1|
L|122937813|gb|ABM68635.1|
L|1817723|gb|AAB42155.1|
L|71916307|gb|AAZ56209.1|
L|157681096|gb|ABV62240.1|
L|2897802|dbj|BAA24918.1|
L|144808|gb|AAA20892.1|
L|125712789|gb|ABN51281.1|
L|7208809|emb|CAB76932.1|
L|316941198|gb|ADU75232.1|
L|353441419|gb|AEQ94264.1|
L|219544539|gb|ACL26277.1|
L|451784663|gb|AGF55631.1|
L|372100184|gb|AEX68682.1|
L|141603871|gb|ABO88214.1|
L|125713488|gb|ABN51980.1|
L|316940505|gb|ADU74539.1|
L|359824960|gb|AEV67733.1|
L|99644508|emb|CAK22316.1|
L|3600052|gb|AAC35539.1|
L|4850284|emb|CAB43040.1|
L|7267803|emb|CAB81206.1|
L|22136600|gb|AAM91619.1|
L|30681638|ref|NP_192843.2|
L|332657566|gb|AEE82966.1|
L|125713287|gb|ABN51779.1|
L|316940717|gb|ADU74751.1|
L|581006|emb|CAA43035.1|
L|334812465|gb|AEH04391.1|
L|6179388|emb|CAB59900.1|
L|1655545|emb|CAA65828.1|
```

CAZy

Accessions

NCBI

GH9 HMM

HMMSEARCH

Known GH9 Protein FASTA
sequences from NCBI

2092

JGI
>1000
metagenomes

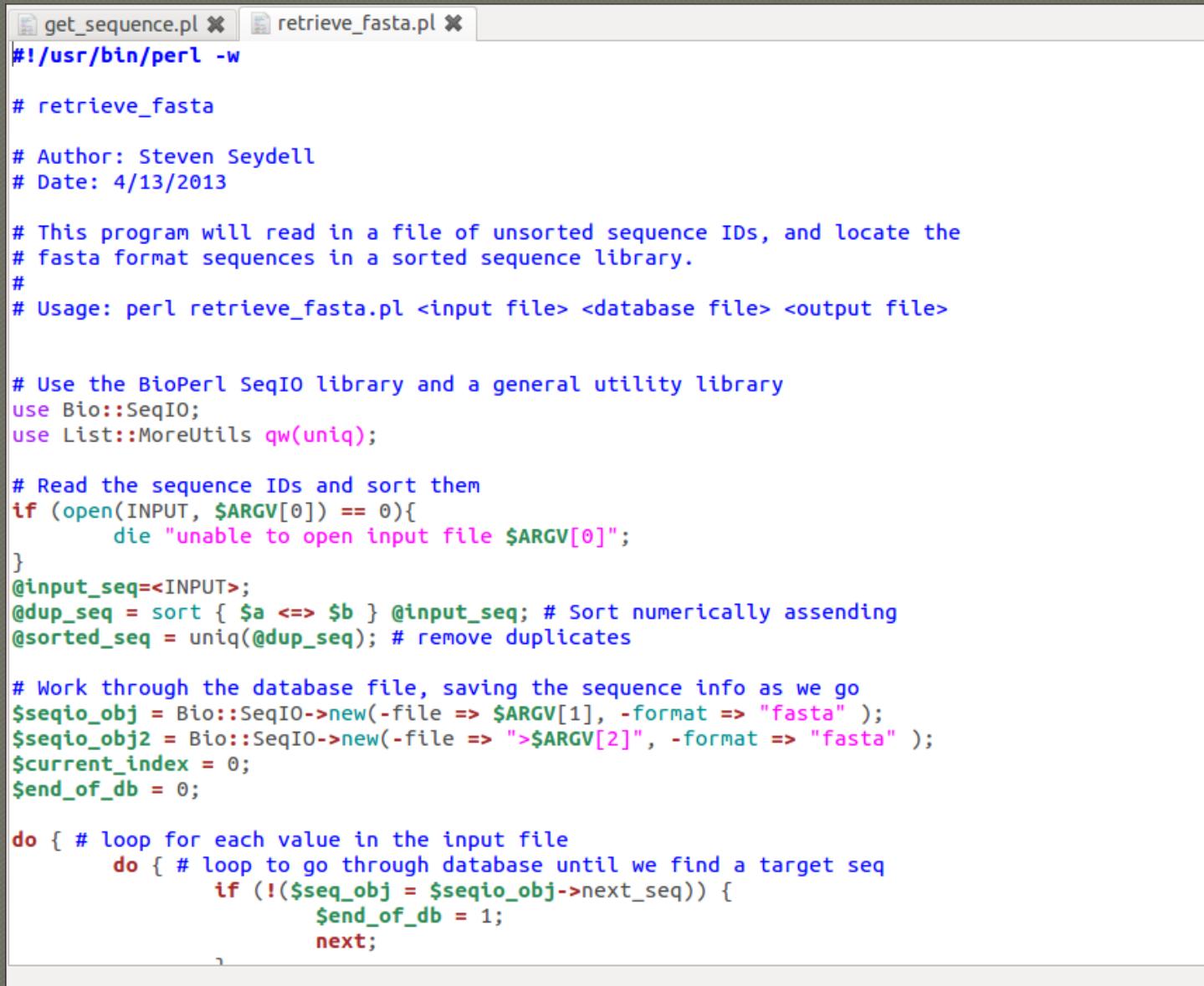
Domains
Positions
1452

Domains
Positions
1777

Domains
FASTA

- Filter
- Extract sequences
- Extract domains

Perl script to extract full length protein FASTA sequences



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

# retrieve_fasta

# Author: Steven Seydell
# Date: 4/13/2013

# This program will read in a file of unsorted sequence IDs, and locate the
# fasta format sequences in a sorted sequence library.
#
# Usage: perl retrieve_fasta.pl <input file> <database file> <output file>

# Use the BioPerl SeqIO library and a general utility library
use Bio::SeqIO;
use List::MoreUtils qw(uniq);

# Read the sequence IDs and sort them
if (open(INPUT, $ARGV[0]) == 0){
    die "unable to open input file $ARGV[0]";
}
@input_seq=<INPUT>;
@dup_seq = sort { $a <=> $b } @input_seq; # Sort numerically assending
@sorted_seq = uniq(@dup_seq); # remove duplicates

# Work through the database file, saving the sequence info as we go
$seqio_obj = Bio::SeqIO->new(-file => $ARGV[1], -format => "fasta" );
$seqio_obj2 = Bio::SeqIO->new(-file => ">$ARGV[2]", -format => "fasta" );
$current_index = 0;
$end_of_db = 0;

do { # loop for each value in the input file
    do { # loop to go through database until we find a target seq
        if (!$seq_obj = $seqio_obj->next_seq) {
            $end_of_db = 1;
            next;
        }
        if ($sorted_seq[$current_index] eq $seq_obj->id) {
            $seqio_obj2->write($seq_obj);
            $current_index++;
        }
    }
} while !$end_of_db;
```

Perl script continued

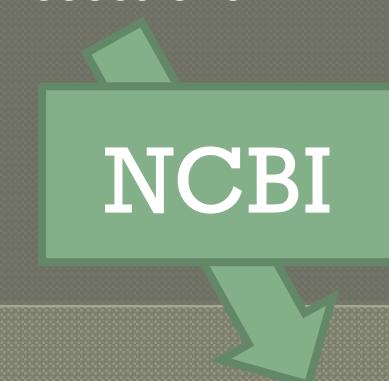
```
        }
    } until ($seq_obj->display_id() == $sorted_seq[$current_index]) || $end_of_db;
next if $end_of_db;

# print the sequence out to the output file
$seqio_obj2->write_seq($seq_obj);
$current_index++;

} until ($current_index == $#sorted_seq + 1) || $end_of_db;
if ($end_of_db) {
    print "ERROR: end of database reached without finding all data.\n";
}

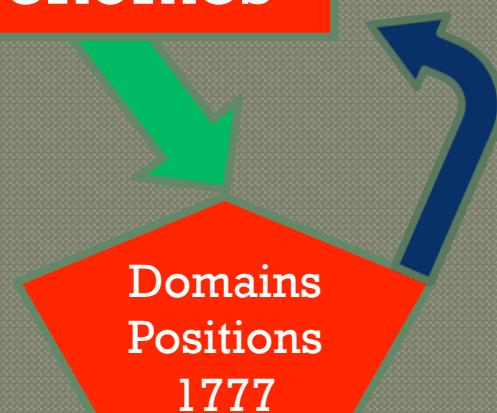
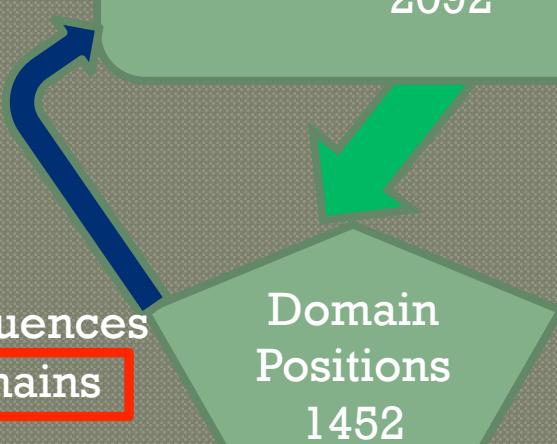
close INPUT;
```

CAZy
Accessions



Known GH9 Protein FASTA
sequences from NCBI
2092

JGI
>1000
metagenomes



- Filter
- Extract sequences
- Extract domains

Python script to retrieve domain regions while simultaneously creating a color code file

```
import re
from Bio import SeqIO

#Parsing FASTA into Biopython
FASTA= SeqIO.parse('GH9.jgi.hmm.out.4.1.1.fa', 'fasta')

#Opening HMM output
hitid = open('GH9.jgi.hmm.out', 'r')

#These are the files that will be written to
o = open('ncbi.output', 'a')
c = open('colors.output', 'a')

#Creating a list for all of the numbers
bank = []
#Creating a list for all rows in HMM output
bigfile = []

#Filling these lists.
for hit in hitid:
    bigfile.append(hit)
    preacc = hit.find(' ')
    accession = hit[:preacc]
    bank.append(accession)

#Actually getting sequences from the FASTA file
for sequence in FASTA:
    #This step is to figure out if the FASTA id
    #is in the HMM output
    if sequence.id in bank:
        if bank.count(sequence.id)>1:
            ind = bank.index(sequence.id)
            cnt=bank.count(sequence.id)
            add =[]
            while cnt !=0:
                add.append(ind)
                ind+=1
                cnt-=1
            for i in add:
                c.write(str(i))
                c.write(' ')
            c.write('\n')
        else:
            o.write(sequence.id)
            o.write('\n')
            o.write(str(sequence.seq))
            o.write('\n')
```

Python cont.

```
        end = ''.join(preend)
        end = int(end)

        print '>', sequence.id, '\n'
        o.write('>%s|%s|%s|\n' % (sequence.id, start+1, end+1))
        c.write('%s|%s|%s|\t#FF0000\n' % (sequence.id, start+1, end+1))
        print sequence.seq[start+1:end+1], '\n'

        o.write('%s\n' % sequence.seq[start+1:end+1])
        ###the +1 was added to offset the Python indices starting at zero.
        inds = []

else:
    ind = bank.index(sequence.id)
    useline = bigfile[ind]
    prestart = re.findall('[0-9]', useline[153:158])
    start = ''.join(prestart)
    start = int(start)

    preend = re.findall('[0-9]', useline[158:164])
    end = ''.join(preend)
    end = int(end)

    print '>', sequence.id, '\n'
    o.write('>%s|%s|%s|\n' % (sequence.id, start+1, end+1))
    c.write('%s|%s|%s|\t#FF0000\n' % (sequence.id, start+1, end+1))
    print sequence.seq[start+1:end+1], '\n'
    o.write('%s\n' % sequence.seq[start+1:end+1])

hitid.close()
o.close()
c.close()
```

CAZy
Accessions



GH9 HMM

HMMSEARCH

Known GH9 Protein FASTA
sequences from NCBI
2092

JGI
>1000
metagenomes

Domain
Positions
1452

Domains
FASTA

Domains
Positions
1777

- Filter
- Extract sequences
- Extract domains

All Domains in FASTA format

```
>gi|20417|emb|CAA42569.1||31|486|
DALEKSILFFEGQRSGKLPTNQRLTWRGDSGLSDGSSYHVDLVGGYYDAGDNLKFGGLPMAFTTMLAWGIIEI
>gi|40672|emb|CAA28255.1||140|570|
DAFKTAMLGMYLLRCGTSVSATYNGIHYSHGPCHTNDAYLDYINGQHTKKDSTKGWHDAGDYNKYVVNAGITV
>gi|45504|emb|CAA31082.1||126|596|
AAFYDALKYFYHNRSGIAIETPYTGGGRGSYASHSRWSRPAGHLNQGANAKDMNVPCWSGTCNYSLNVTKGW
>gi|144416|gb|AAA23086.1||39|473|
AEALQKSMFFYQAQRSGDLPADFPVSWRGDSGLTDGADVGKDLTGGWYDAGDHVKFGFPMAFSATMLAWGAIE
>gi|144808|gb|AAA20892.1||78|510|
GEALQKAIFFYECQRSGKLDPLSTLRLNWRGDSGLDDGKDAGIDLTGGWYDAGDHVKFNLPMSYSAAMLGWAV
>gi|166947|gb|AAA32912.1||31|486|
DALEKSILFFEGQRSGKLPTNQRLTWRGDSGLSDGSSYHVDLVGGYYDAGDNLKFGGLPMAFTTMLAWGIIEI
>gi|167883|gb|AAA52077.1||27|452|
CSLLENALMFYKMNRRAGRLPDNDIPWRGNSALNDASPNSAKDANGDGNLSGGYFDAGDGVKFGLPMAYSMTMI
>gi|296791|emb|CAA39010.1||32|462|
GEALQKAIFMFEFQRSGKLPEKNRDNRGDGSGLNDGADVGKDLTGGWYDAGDHVKFNLPMAYSQTMLAWAAYE
>gi|305112|gb|AAA68129.1||103|551|
SNIKSFYYQRSGVELPERLAGIWARPAAHLDKLEFHPTMERAGLWNAHGGWYDAGDYGKYIVNGGVSVATLM
>gi|310897|gb|AAC06387.1||278|746|
ELRVDALSFYYPQRSGIEILDIAPIGYGRPAGHIGVPPNQGDTDVPCAPGTCDYSLDVSGGWYDAGDHGKYV
>gi|349601|gb|AAA02563.1||41|489|
ADALAKAILFFEGQRSGKLPSQRVKWRDSDLSDGKLQNVNLMGYYDAGDNVKFGWPMAFSTSLLSWAAVE
>gi|530014|emb|CAA56918.1| 309 811
QMKYDALAFFYHKRSGIPIEMPYAGGEQWTRPAGHIGIEPNKGDTNVPTWPQDDEYAGIPQKNYTKDVTGGW
>gi|530014|emb|CAA56918.1| 309 811
```

Domains
FASTA

JGI: 1777 GH9 domains from 211 metagenomes
NCBI 1452 GH9 domains

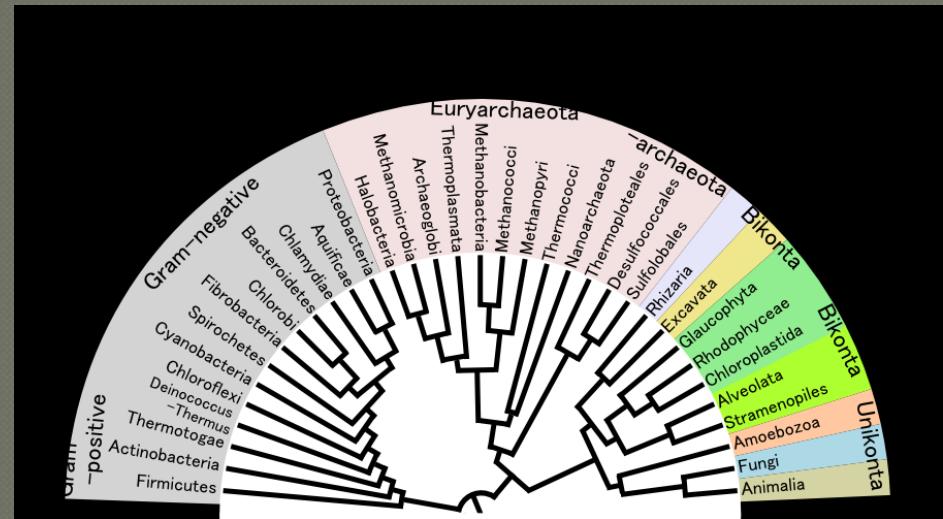
MAFFT

Domains
Aligned

FastTree

ITOL

Phylogeny

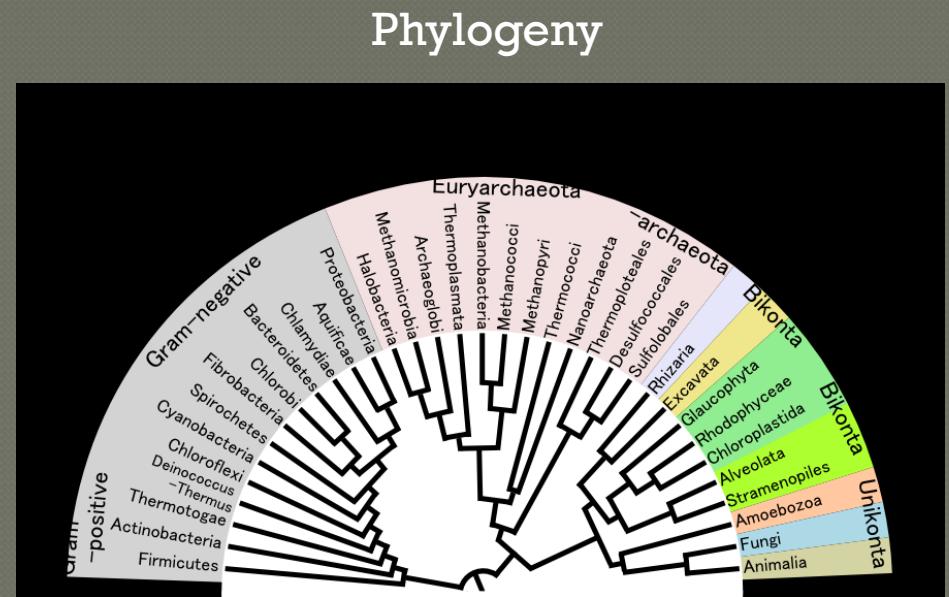
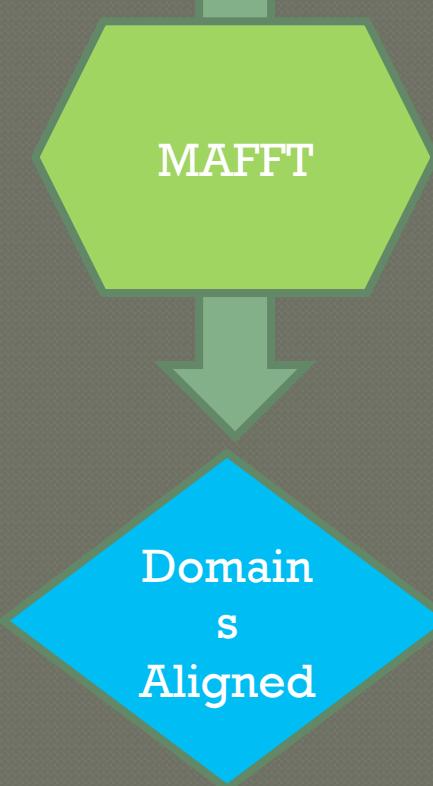


mafft –auto totalregions.fasta > totalregions.fasta.aln

```
gi|20417|emb|CAA42569.1||31|486|
-----DALEKSILFFEGQR-
-----SGKL-P-TN-----QR-LTWR-----GDS
LS-DGS-----S-----Y-----
-----H--VDLVGGYYDA-GDNL--KFGLPMAFT-----
TTMLAWGIIEF-----GCL-----MP-----E-----
-----QV-----EN-----
AR-AALRWST---DYLLK-AS-----TATS-----NSLYV
-----VGE--PN-----A-----DHRC--WER-----PED
DT-----PRN--VYKVS-T-QNP-
G-S-----DVAEAETAAALAAAISIVF--G-----D
-----S-----DSS-----Y-----STK-LL-----HT-----
-----AV--KVFE-----FA-----DQ-----
-----YRG-----SYS-DS--LG--S-VVCP--FY-----CSYS-
-----G-YN-----DE--LL-W-----G-----
-----AS-----W-LH-----RA-----S-----
-----Q-N-A-SY-MTY-----I-----
-----Q-SNGHTLGA-----DDDDYSFSWDDKRV-----GT
-----K-----VL-----
-----LS-----
-----KGFL-----Q-----D-----RI-----
-----EELQL--YKVHTD--NYICSLI-----P-----
TSSFQ--A-Q-Y-TPG-----GLLY-----KG-----
-----SA--S--N--LQYV-----T-STA-----
LLLTYA-N-YLN-----S-SG--G-H-----ASCG--TT-----TVTA-
KN-LI-----SLAK-----KQVDYILGQN-----PA-KM-----SYMVGF-----G
E-----RY-----PQHVHHRG-----SSLPSVQVHPNSIPCNAG-----FQ-
LY-----SS-----PPNPNIL-----VGAILGG-----P-D-----
-----N-R-----
-----D-----
-----SFS-D--DRN-----N-----YQQSEPATYINAPLVG
LAF-----
gi|40672|emb|CAA28255.1||140|570|
-----DAFKTAMLGYLLR-
-----CGTSV-----SA-TYNGIHSHGPCHT--N
AYLDYI-----NGQHTK-----
-----KDSTKGWHDA-GDYN--KYVVNAGIT-----
-----VGSMFLAWEHF-----KDQ-----LE-----P-----
-----VAL-EI-----PE-K-----NNS-----IPD-----
FL-DELKYEI--DWILT-MQYP-DG-----SGRVAH
-----VS-----TRNF-GG-----FIM-PEN
-----EHDE-----RFFVP-----WS-----
-----S-----AATADFVAMTAMAARIF-----R
Y-----DPQ-----Y-----AEK-CI-----NA-----
-----AK-VSYE-----FL--K-NN-----PAN-
-----VFA-----N-QSGFSTG--EY-----AT-----
```

Domains
FASTA

JGI: 1777 GH9 domains from 211 metagenomes
NCBI 1452 GH9 domains



FastTree

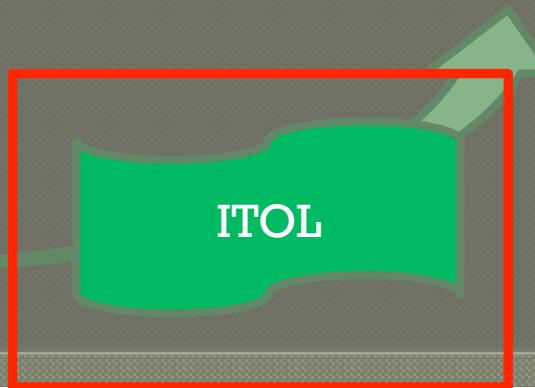
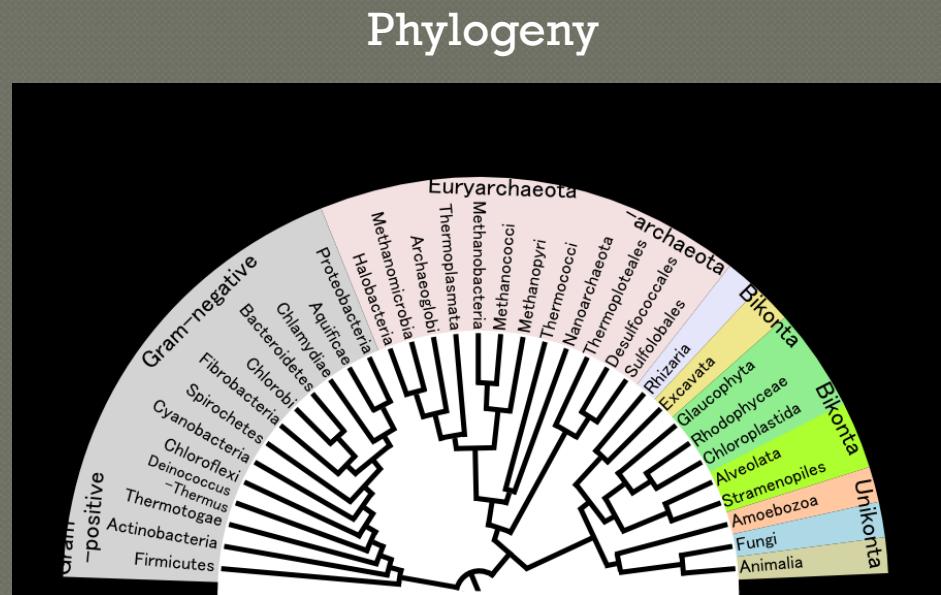
ITOL

/home/mrupani/Downloads/Fastree totalregions.fasta.aln> treeproject1.nwk

```
(gt|361075211|gd|AEW10553.1||158|594|:0.82921,((gt|16189816|gd|ABX6847.1||2|45|:0.49742,gt|328449638|gd|AEB15354.1||290|715|:0.65377)0.837:0.15552,((2096726972|130|0|:0.66222,(gi|306533049|gb|ADN02583.1||295|727|:0.02633,gi|339412399|gb|AEJ61964.1||297|729|:0.03661)1.000:0.44226)0.883:0.16893,((2021679479|94|511|:0.84112,(700738251|93|528|:0.00194,(7039424245|93|528|:0.0,7068761396|93|528|:0.0):0.00606)1.000:0.66853)0.964:0.16630,((gi|291519188|emb|CBK74409.1||3|422|:0.53313,((2098892475|309|743|0.2165244832|309|743|:0.0,2197419208|309|743|:0.0):0.45087,((7035930076|378|818|:0.0,7040585422|279|719|:0.0):0.00486,((7003238918|378|818|:0.0,7018916353|378|818|:0.00014,7058730133|2|385|:0.00276)0.930:0.00015)1.000:0.28741)0.602:0.07505)0.847:0.07548,((gi|317022228|gb|ADU86915.1||323|781|:0.33161,gi|322512572|gb|ADX05690.1||334|73|:0.44653)0.999:0.22556,((2018710611|327|757|:0.00015,7071797159|327|757|:0.000257)0.868:0.00514)1.000:0.12487,((gi|295094191|emb|CBK8328|1||325|755|:0.0,7009639127|313|743|:0.0,7017083860|288|718|:0.0,7019679862|325|755|:0.0,7030874394|325|755|:0.0,7034580656|325|755|:0.0,7035067304|325|755|:0.0,70430111|325|755|:0.0,7045203053|325|755|:0.0,7051893785|32|462|:0.0,7063391255|325|755|:0.0,7068927468|325|755|:0.0):0.00015,7004312348|320|750|:0.00768)0.995:0.09094)1.000:0.769,((7007803849|121|553|:0.12588,((7005508281|333|763|:0.0,7032423095|333|763|:0.0,7051284795|330|760|:0.0,7058577811|330|760|:0.0):0.03470,(2018711858|326|756|:0.0025((7036459230|326|756|:0.0,7074543537|326|756|:0.0):0.00014,(7024853441|326|756|:0.0,7056888295|313|743|:0.0):0.00258)0.000:0.00014)0.936:0.02349)0.998:0.09543)0.994:0.156,(7042795983|347|780|:0.00243,((7027723249|34|467|:0.0,7044872196|347|780|:0.0,7068093337|347|780|:0.0):0.00014,(((7029760161|32|465|:0.0,7078774501|347|780|:0.0):0.251,(7034569998|347|780|:0.0,7046504226|346|779|:0.0,7068709864|347|780|:0.0,7074625324|300|733|:0.0):0.00248)0.739:0.00014,(7003566266|135|568|:0.0,7004303494|347|780|:0.0,7008497395|347|780|:0.0):0.00014)0.00005:0.00015,7045792230|3|368|:0.00015)0.916:0.00252)1.000:0.00016)0.49738)0.540:0.08041)0.969:0.13572)0.985:0.13415)0.994:0.764,((7016113225|287|715|:0.34609,((gi|291543673|emb|CBL16782.1||291|717|:0.17640,(7028711724|292|645|:0.06061,7036159898|291|716|:0.06670)0.971:0.06508)0.999:0.13610,(7007287208|289|714|:0.00396,7058485163|2|369|:0.00164)1.000:0.21378,(7046642831|19|448|:0.31349,(7016108178|261|688|:0.21807,(7025046230|189|553|:0.00016,7056890015|288|2|:0.00014)1.000:0.19349)0.066:0.02697)0.935:0.05664)0.848:0.05927)0.050:0.06456)1.000:0.20790,(((2197437418|3|363|:0.20540,((gi|302577215|gb|ADL51227.1||287|719|:0.062|,gi|302577216|gb|ADL51228.1||287|719|:0.04867)1.000:0.26507,((gi|49425363|gb|AAT66046.1||288|720|:0.0,gi|302577217|gb|ADL51229.1||288|720|:0.0):0.17399,gi|451784938|gb|F55906.1||291|723|:0.16514)0.161:0.03821)0.210:0.02147),(gi|280977773|gb|ACZ98604.1||290|721|:0.34949,gi|326539941|gb|ADZ81800.1||155|590|:0.48750)0.078:0.894)0.947:0.06256,gi|326540549|gb|ADZ82408.1||291|722|:0.29686)1.000:0.19524)0.990:0.12511)0.810:0.07680)0.885:0.06195)0.051:0.03806)0.977:0.12511)0.868:0.06546,(((70061338|90|495|:0.00016,(7032446787|90|543|:0.00345,7058211306|2|394|:0.00268)0.507:0.01202)1.000:0.52513,((7003308015|101|553|:0.00014,7020937639|101|514|:0.00014)0.705)0.00014,(7040034874|101|553|:0.00488,(7036289420|114|566|:0.00484,((7001367146|101|553|:0.0,7050067964|101|553|:0.0):0.00242,(gi|290770192|gb|ADD61951.1||101|553|:0.00247068759905|101|553|:0.00014)0.000:0.00014)0.000:0.00016)0.763:0.00241)0.872:0.00723)1.000:0.48539)0.944:0.22118,(gi|580767|emb|CAA39264.1||90|543|:0.00076,gi|302395425|b|ADL34330.1||90|543|:0.00496)1.000:0.81533,((7067743183|106|550|:0.00506,(7018947786|106|550|:0.00254,7059347542|106|488|:0.00289)0.779:0.00253)0.320:0.00015,(7007295|0|106|550|:0.00014,7071847863|106|550|:0.00251)0.880:0.00245)1.000:0.46128,(7058580318|105|543|:0.00014,((7018924293|4|434|:0.0,7045810953|4|434|:0.0):0.00014,70490917|2|438|:0.00014)0.951:0.00757,((7016817292|105|543|:0.01520,7026779141|105|543|:0.01015)0.764:0.00244,((7009804907|105|543|:0.0,7027060507|105|543|:0.0):0.00498,7017346|5|105|543|:0.00250)0.786:0.00250)0.758:0.00247)0.762:0.00496)1.000:0.33879)0.959:0.18715)0.935:0.10414)0.962:0.12926,(gi|158451923|gb|ABW39322.1||96|540|:0.12633,(gi|1451925|gb|ABW39323.1||96|498|:0.04245,gi|158451927|gb|ABW39324.1||96|540|:0.09465)1.000:0.15902)0.995:0.36791,((7018777507|80|517|:0.0,7049082596|80|517|:0.0):0.00252,7026819876|80|517|:0.0,7027064309|80|517|:0.0):0.00014)0.995:0.00015,(7058512564|80|517|:0.00251,(7016873445|80|517|:0.0,7017359399|80|517|:0.0):0.00251)0.000:0.00012,709822900|80|517|:0.00014)0.849:0.00247)1.000:1.51584)0.173:0.17002)0.995:0.19583,((2120645395|163|513|:0.08432,(gi|219999967|gb|ACL76568.1||145|583|:0.05737,gi|373945911|b|AEY66832.1||145|583|:0.04764)0.975:0.10333)1.000:0.63931,(2078117575|137|491|:0.39239,((2200954926|100|506|:0.54653,(2061989454|15|397|:0.29403,(gi|296837349|gb|ADH593.1||96|536|:0.29590,(gi|339290550|gb|AEJ44660.1||89|530|:0.07249,((gi|13274207|emb|CAC34051.1||89|530|:0.0,gi|257479162|gb|ACV59481.1||89|530|:0.0,2165281811|108|549|:0.0):0.00014,2120726710|114|502|:0.00015)0.951:0.06465)1.000:0.66185)0.994:0.22974)0.874:0.07758)0.703:0.04669,(2053635072|95|535|:0.00755,2061979932|95|535|:0.00273)1.00:0.35481,(gi|215512090|gb|ACJ68032.1||96|536|:0.46748,(gi|407726671|dbj|BAM46669.1||94|535|:0.34625,(gi|247543162|gb|ACT00181.1||97|536|:0.16375,2201638846|97|536|:0.111)1.000:0.29852)0.424:0.05060)0.947:0.09556)0.988:0.12230)0.992:0.16609)0.486:0.08222)0.496:0.05918)0.556:0.06661)0.306:0.07179,((7021967666|106|435|:0.53940,((70076413|160|505|:0.0,7019631097|160|505|:0.0,7032448374|160|505|:0.0):0.00933,(7058482925|86|431|:0.00313,(7049091470|2|340|:0.00014,7074627458|59|404|:0.00014)0.313:0.00016)0.20:0.00014)1.000:0.44126)0.986:0.29047,(gi|302396621|gb|ADL35526.1||159|504|:0.46562,(7006736380|138|482|:0.66661,2096690233|50|387|:0.51473)0.870:0.15604)0.964:0.21832)0.000:0.89296,((gi|395811250|gb|AFN73999.1||110|556|:0.51868,(gi|383803219|gb|AFH50299.1||111|557|:0.36154,2137806461|115|561|:0.38574)0.859:0.11262,(gi|335392848|gb|AE7926.1||2|382|:0.33989,gi|335392858|gb|AEH57931.1||2|381|:0.49535)0.933:0.12389)0.731:0.12307)1.000:0.30644,((((2005585746|33|381|:0.71362,(((gi|40672|emb|CAA28255.1||0|57|:0.0,gi|125713568|gb|ABN52060.1||140|570|:0.0,gi|316940425|gb|ADU74459.1||140|570|:0.0,2018708216|140|570|:0.0):0.17923,((gi|219998532|gb|ACL75133.1||142|571|:0.2018791884|142|571|:0.0):0.02753,gi|373945140|gb|AEY66061.1||146|571|:0.04585)0.973:0.06382,(gi|373945141|gb|AEY66062.1||132|565|:0.15956,2120625578|142|575|:0.10700)0.33:0.10442)1.000:0.16928)0.990:0.13103,(gi|335392872|gb|AEH57938.1||2|361|:0.45438,(gi|312443557|gb|ADQ79913.1||116|544|:0.33777,(2098960686|87|514|:0.0,2197335972|114|514|:0.0):0.29609)0.976:0.11807)0.992:0.15559)0.935:0.09179)0.537:0.06549,(2201683708|130|550|:0.48882,(2201192472|148|640|:0.57285,2201192480|147|642|:0.50713)1.000:0.560)0.930:0.19003)1.000:0.53017,(gi|335392724|gb|AEH57864.1||2|383|:0.47072,(gi|225793406|gb|AC033496.1||89|535|:0.53734,(gi|335392706|gb|AEH57855.1||2|391|:0.42887,(gi|3392726|gb|AEH57865.1||2|388|:0.37548,(gi|385158833|gb|AFI43954.1||2|393|:0.30432,((gi|335392708|gb|AEH57856.1||2|390|:0.08785,(gi|335392700|gb|AEH57852.1||2|390|:0.13384)0.996,(gi|333494664|gb|AEF56863.1||94|542|:0.00234,2020760641|100|548|:0.00014)1.000:0.19131)0.773:0.0.45559)0.549:0.03045)0.505:0.03622)0.891:0.06165)0.484:0.03618)0.960:0.4908)0.989:0.20604)0.992:0.19188)0.985:0.18253,((((gi|158451913|gb|ABW39317.1||138|573|:0.0,gi|158451921|gb|ABW39321.1||138|573|:0.0):0.04112,gi|158451919|gb|ABW39320|treeproject1.nwk
```

Domains
FASTA

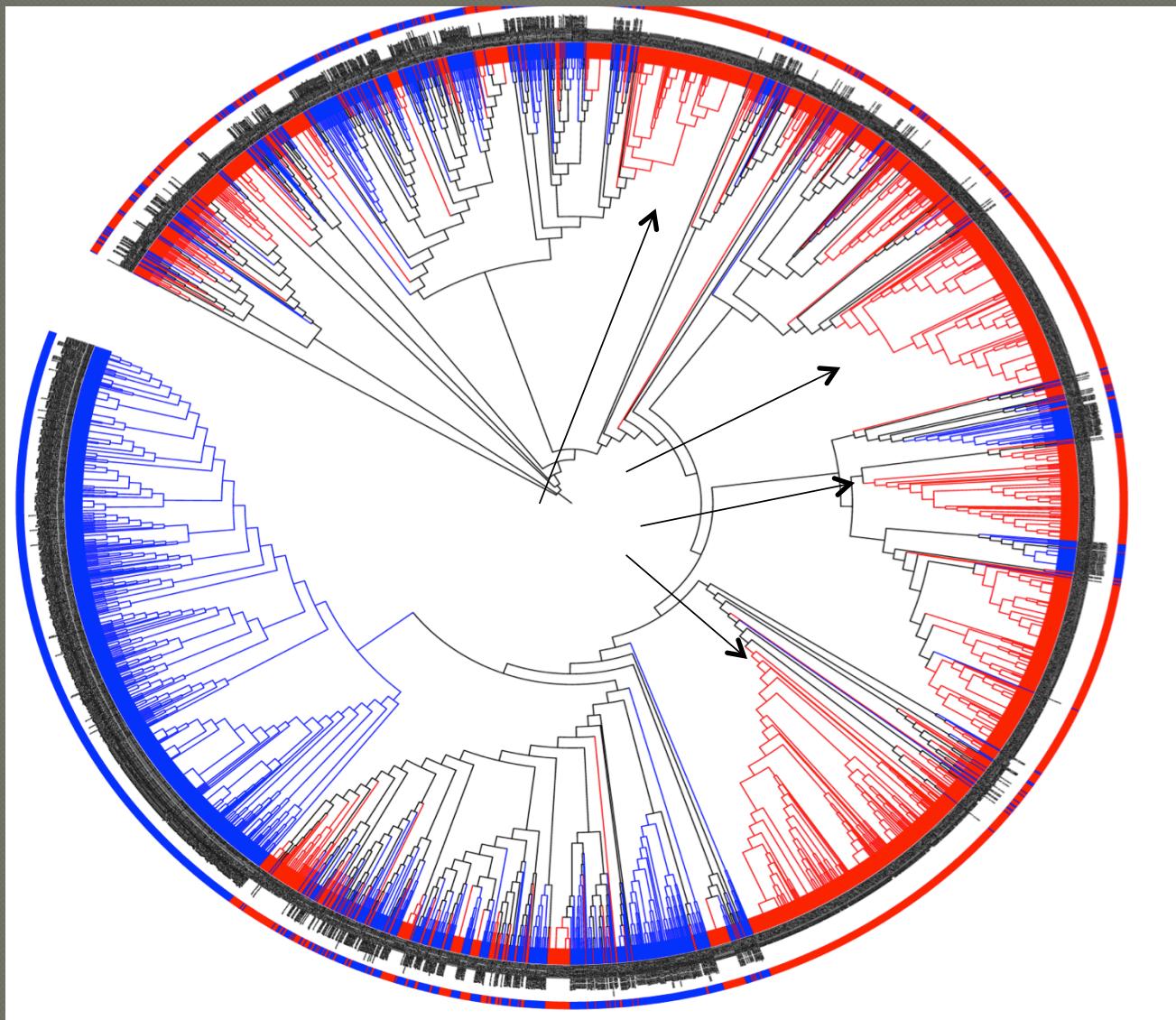
JGI: 1777 GH9 domains from 211 metagenomes
NCBI: 1452 GH9 domains

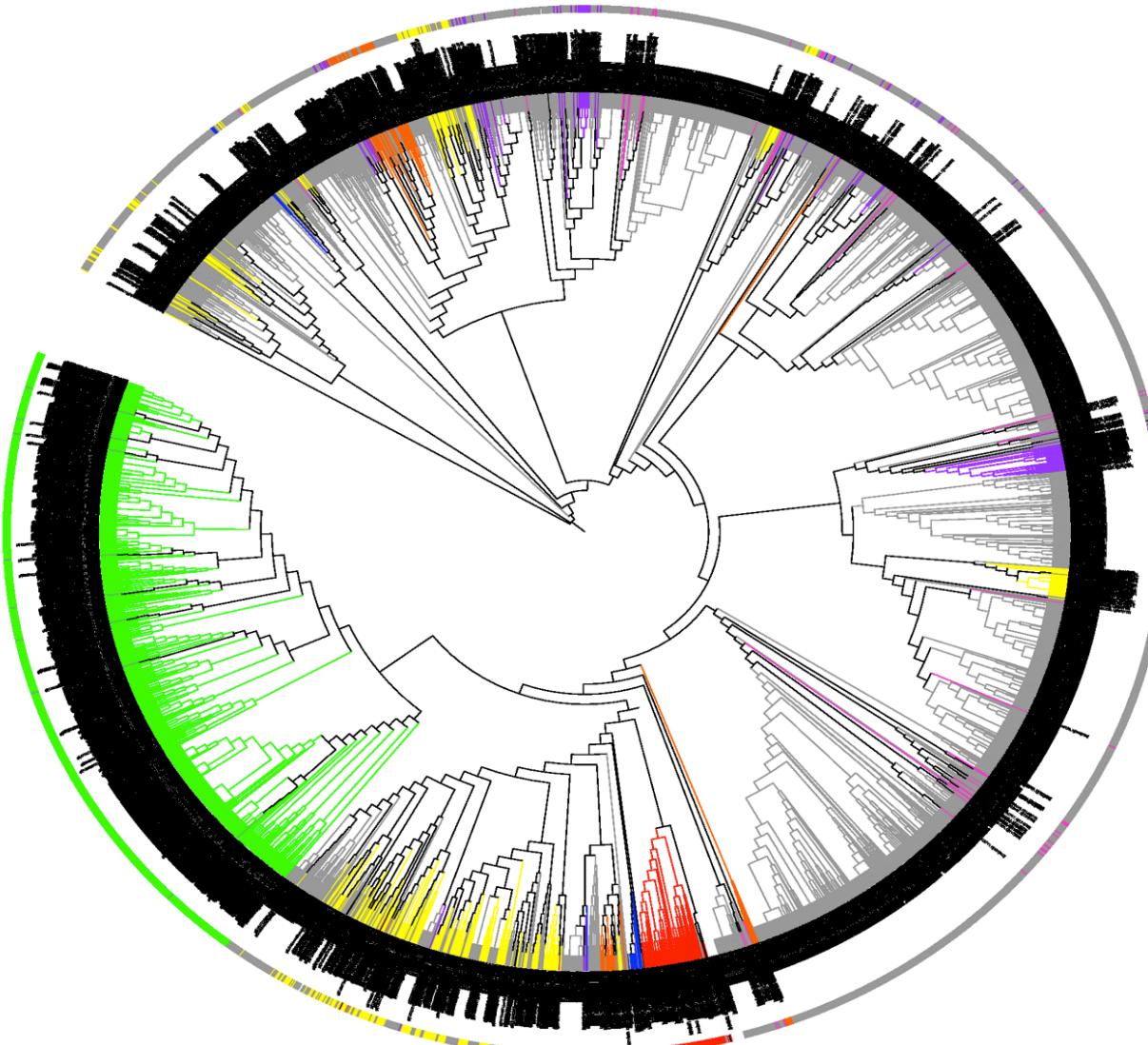


Tree created in iTOL-

Red: JGI

Blue: NCBI





Misc. Bacteria

Fungi

Cyanobacteria

Bacteroidetes

Metazoa

Actinobacteria

Proteobacteria

Viriplante

Firmicutes

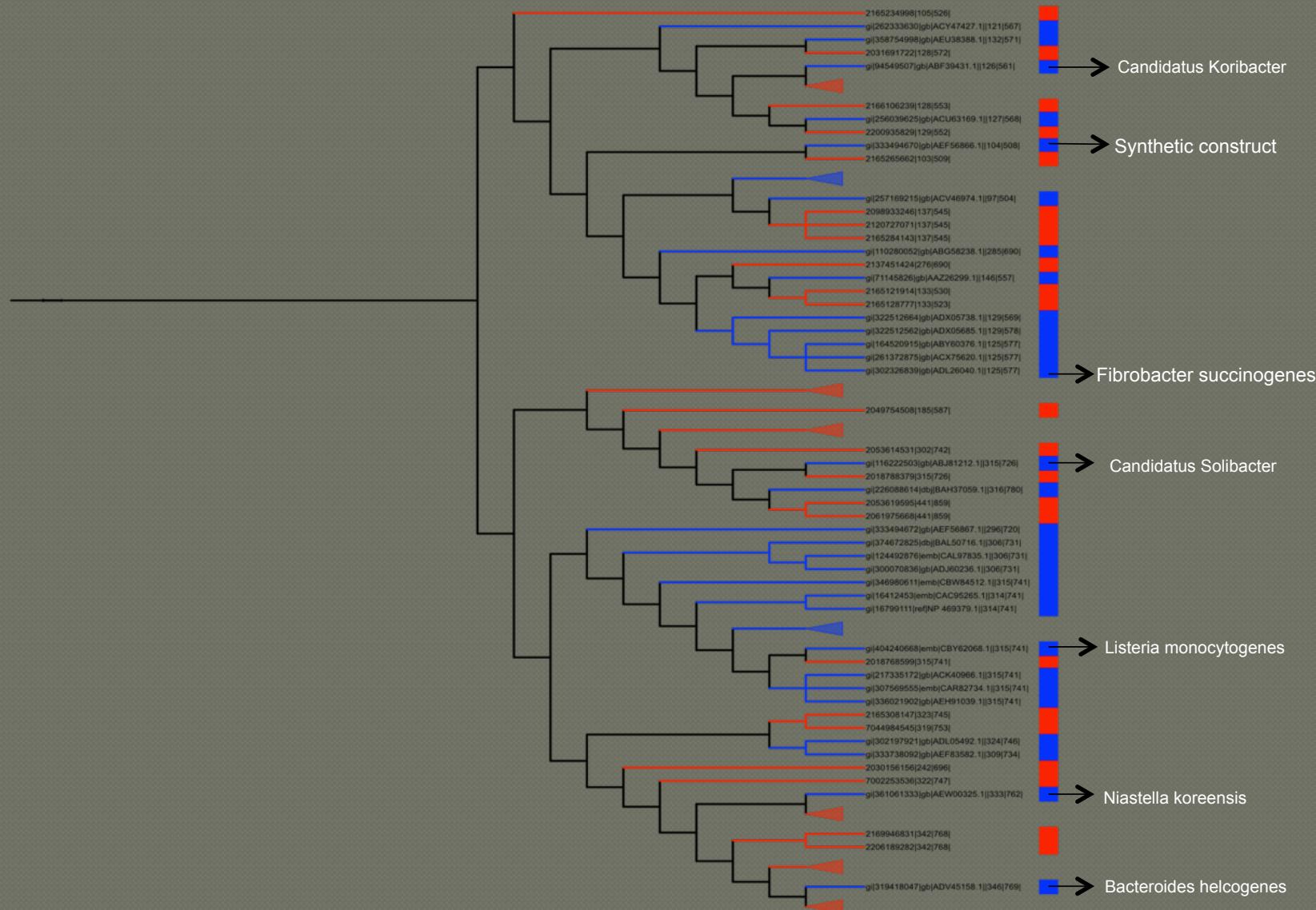
Protists

Stramenopiles

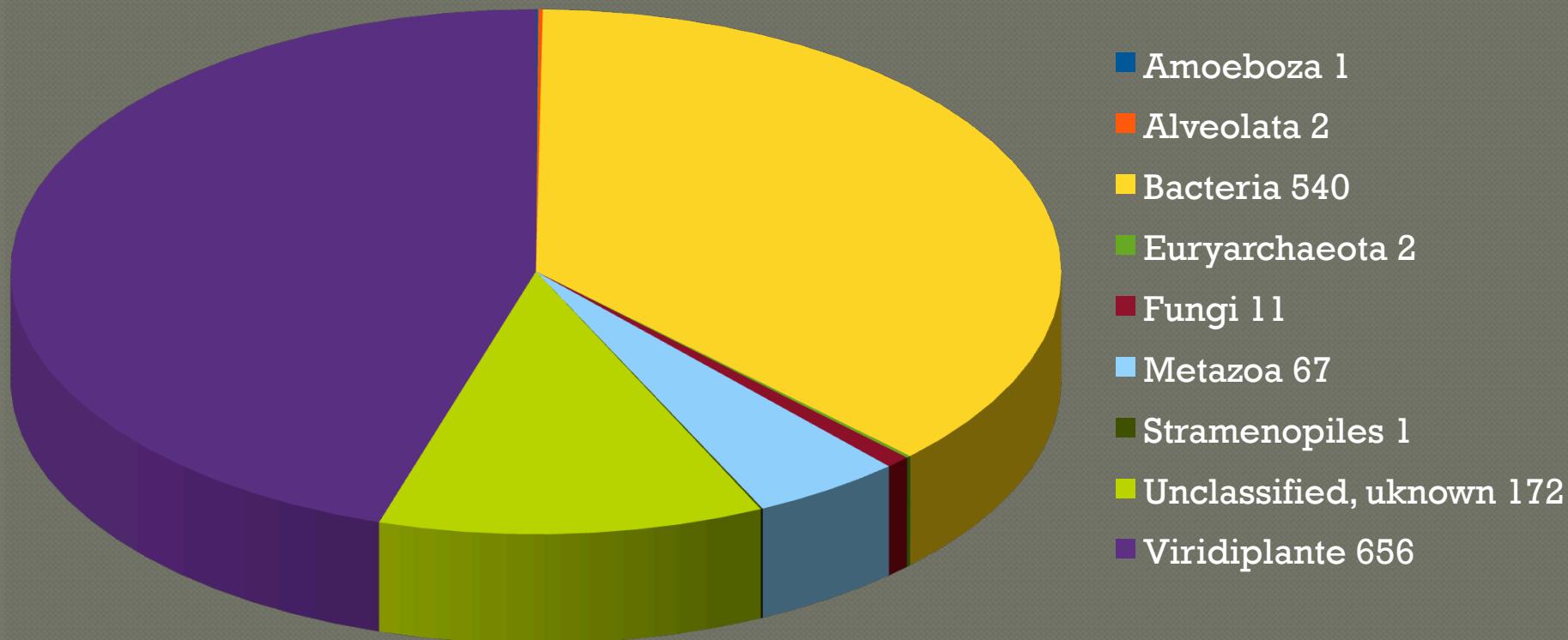
Euryarchaeota

**All unknown
organisms - Gray**

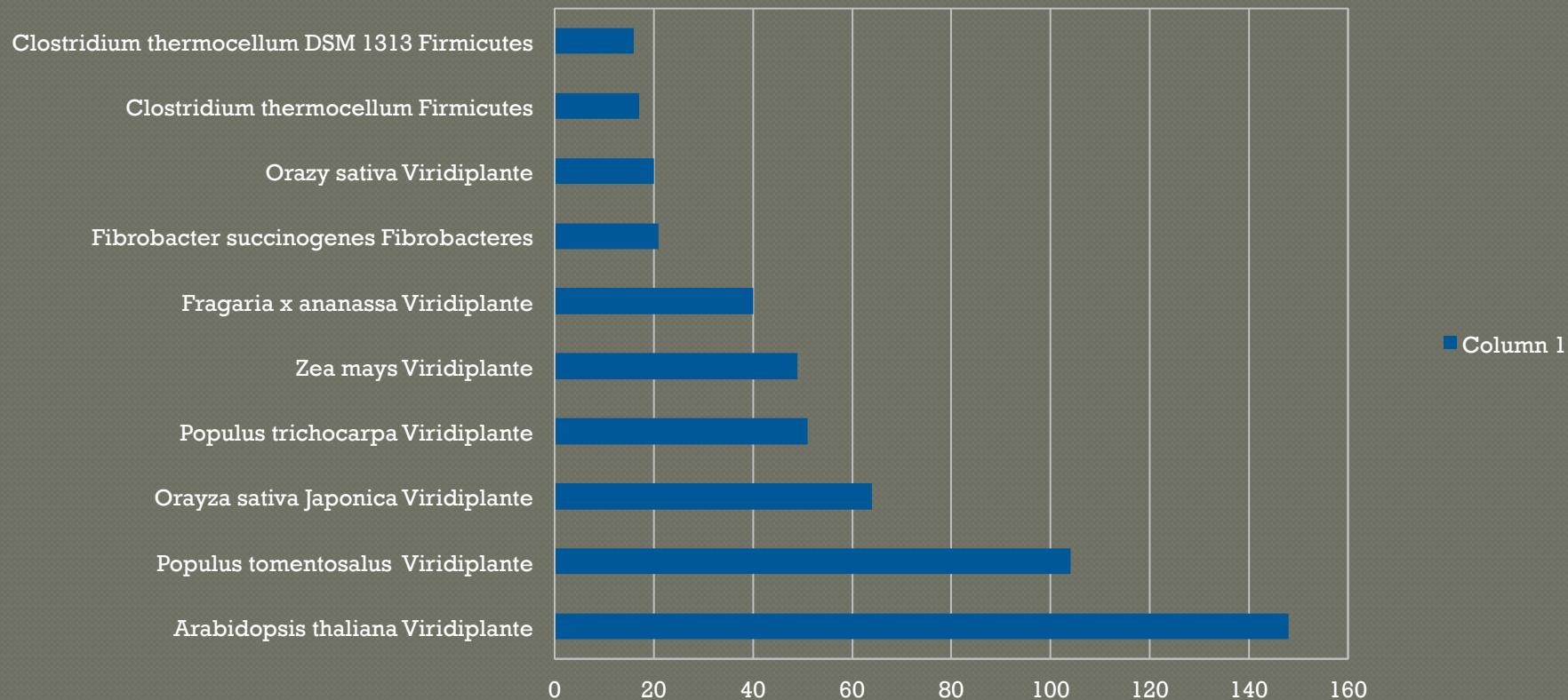
Un-banked GH9 proteins



Phylogenetic distribution of 1452 GH9 sequences in NCBI That contain biochemically confirmed cellulases



Top 10 organisms with multiple GH9 domains



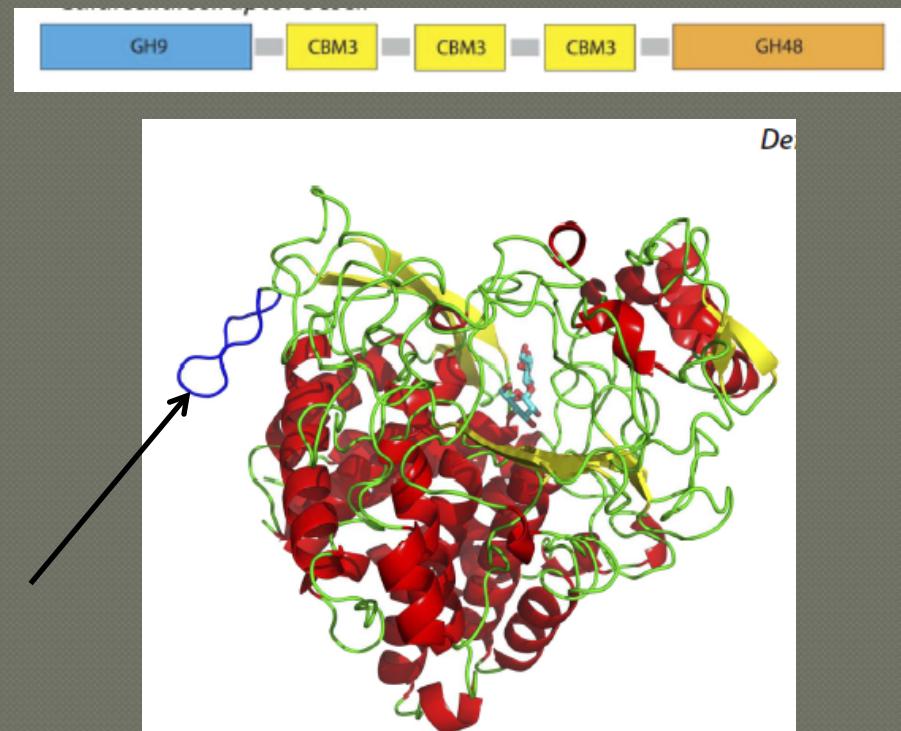
Examples of metagenomic sources of GH9 domains:

- Underneath the Human Tongue
- *Bankia setacea* gills
- Rice straw compost
- Hot springs
- Reindeer rumen
- *Macropus eugenii* forestomach



Novel GH9? What Next?

- Locate, Isolate, Experimentally Test
- Computational identification is not reliable
- Efficiency
 - Synergy with GH48
- Working Conditions
- 3D Structure



Sukharnikov et al., 2012

References

- Aspeborg, H., Coutinho, P.M., Wang, Y., Brumer, H., and B. Henrissat. 2012. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5. *BMC Evolutionary Biology*. **12**: 186-197.
- Finn, R.D., Clements, J., and S.R. Eddy. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research. Web Server Issue* 39:W29-W37.
- Katoh, M and Kuma, M. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**:3059-3066
- Letunic I. and Bork P. 2006. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**(1):127-8
- Price, M.N., Dehal, P.S., and Arkin, A.P. 2009. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26**:1641-1650
- Sukharnikov, L.O., Alahuhta, M., Brunecky, R., Upadhyay, A., Himmel, M.E., Lunin, V.L., and I.B. Zhulin. 2012. Sequence, Structure, and Evolution of Cellulases in Glycoside Hydrolase Family 48. *The Journal of Biological Chemistry*. **287**: 41068-41077.