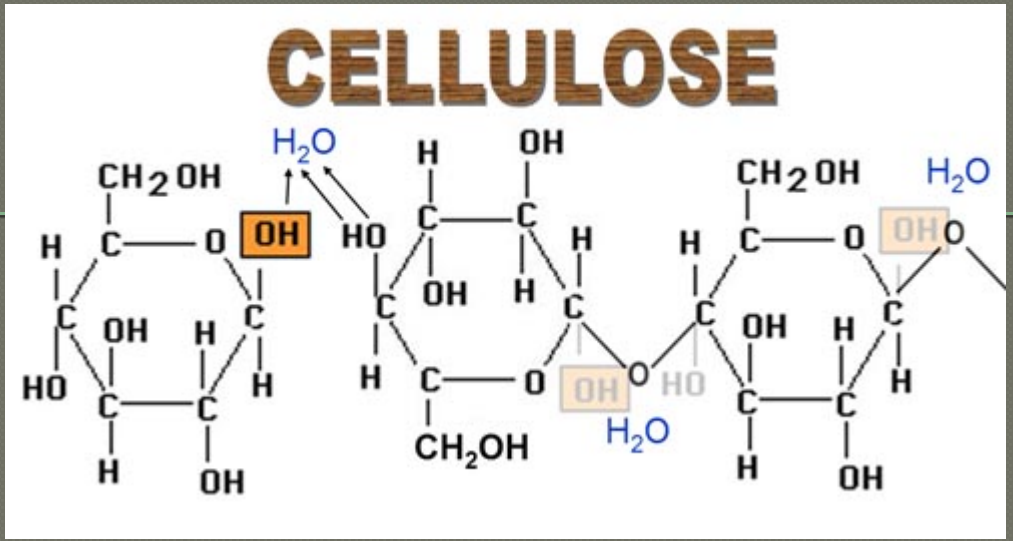
A Search for Novel GH9 Homologues within a Metagenomic Database

Tom Bean
Brenda Pierson
Steve Seydell
Bill Wysocki

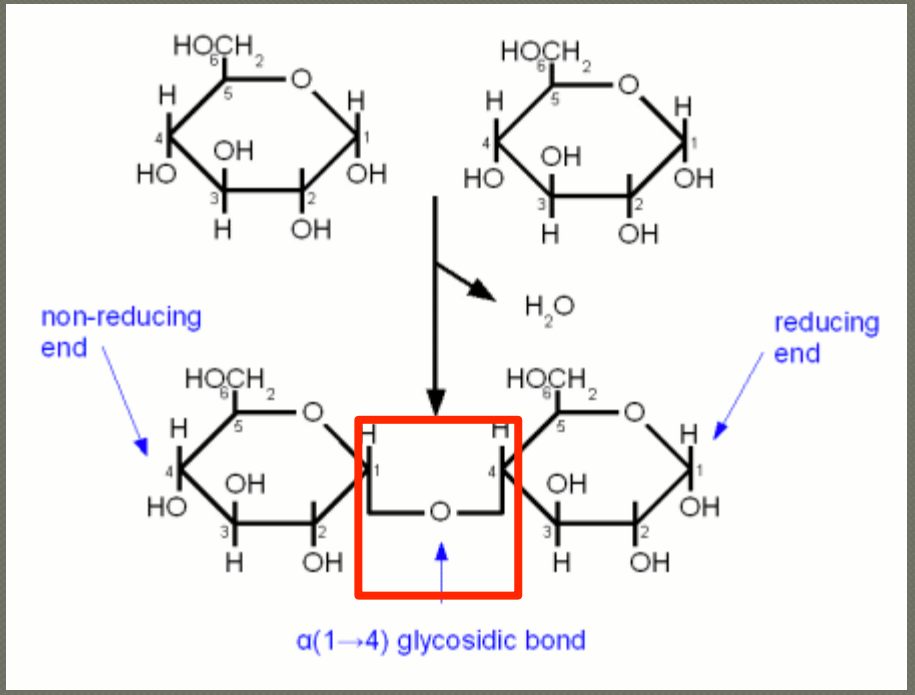


- Polysaccharide found in plant cell walls.
- Very numerous in nature
- Broken down by cellulase
  - Defined by using cellulose as a substrate.

Sukharnikov et al., 2012

Glycoside Hydrolases

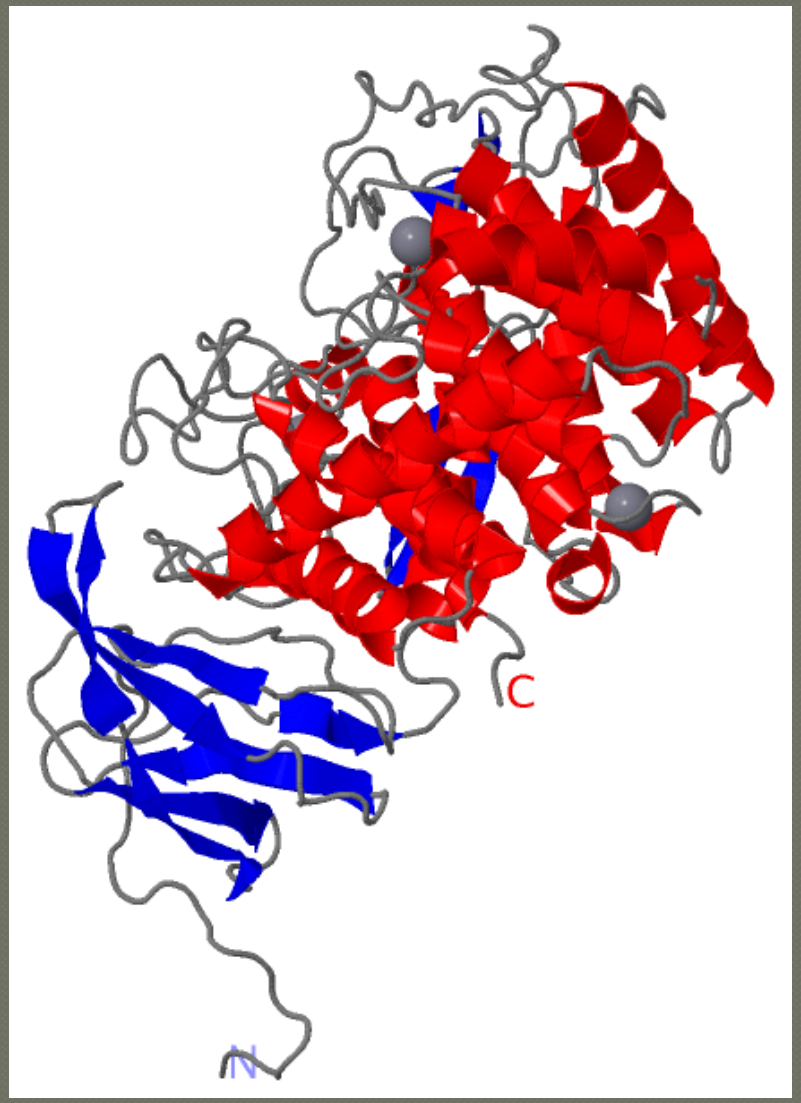
- Break glycosidic linkages
- Understanding is crucial to understanding the ‘CAZome’



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



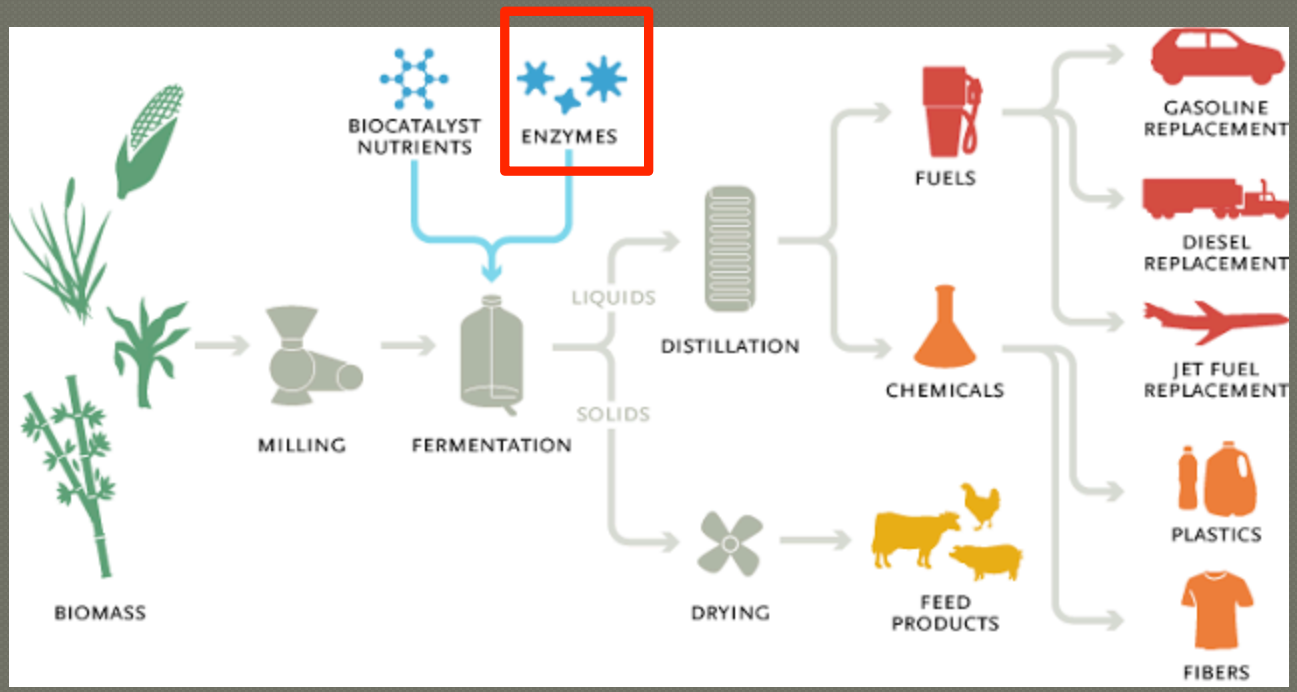
In Nature...

- 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.

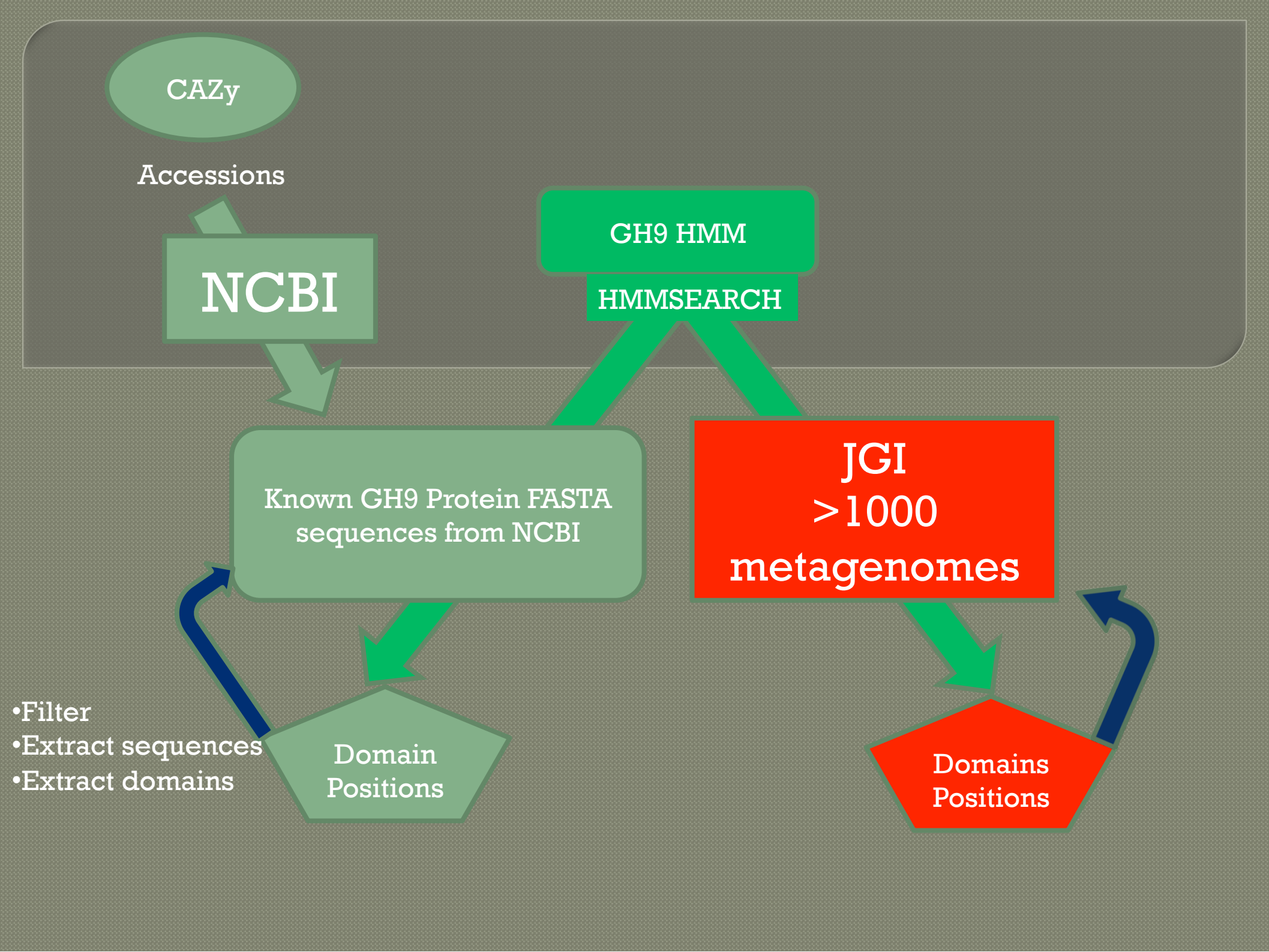


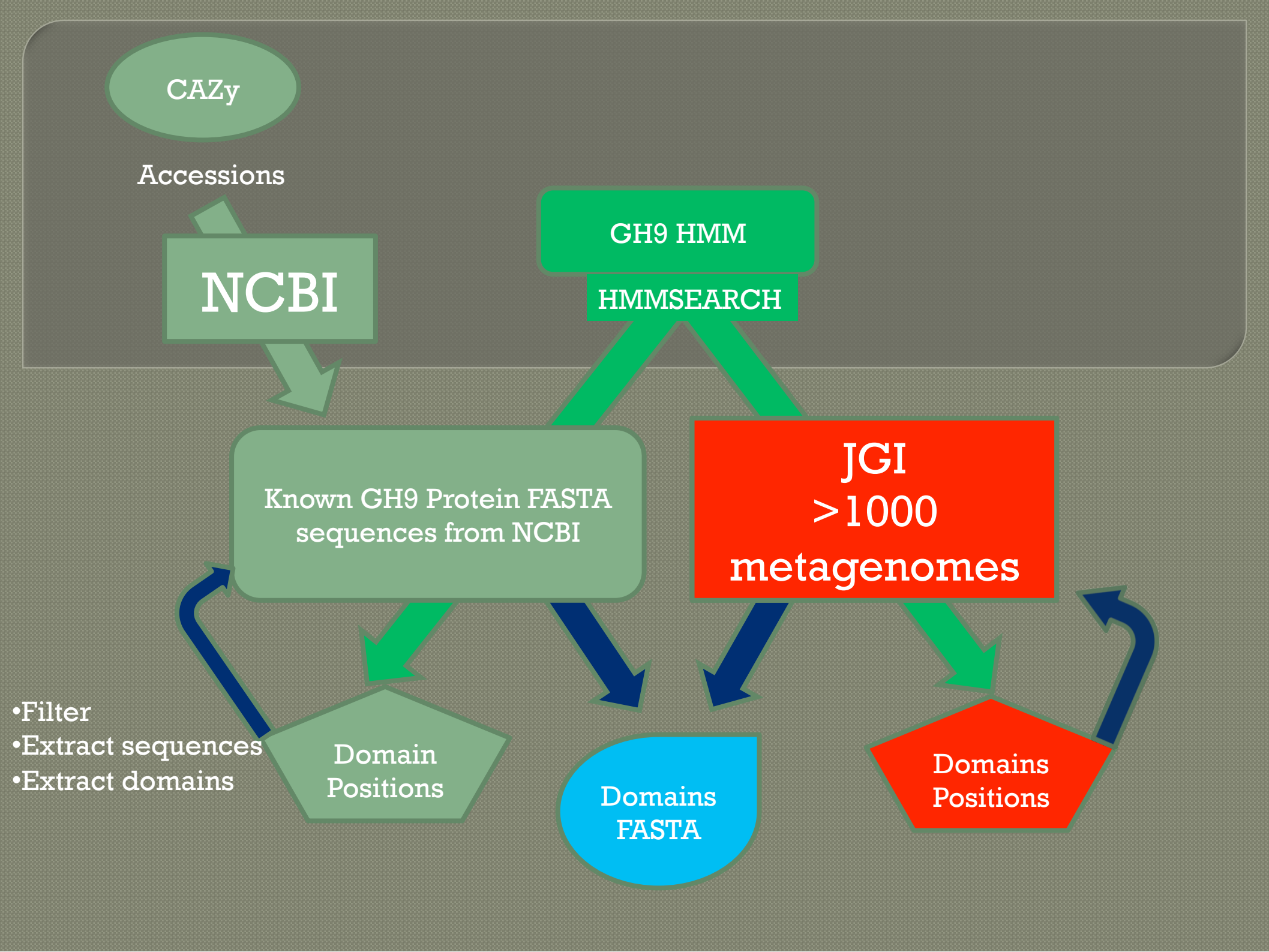
Gevo first ferments isobutanol from biomass (cellulose-derived sugars), then converts the isobutanol to other high value materials.

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
•Filter
•Extract sequences
•Extract domains
Domain Positions
Domains Positions


CAZy
Accessions
NCBI
GH9 HMM
HMMSEARCH
Known GH9 Protein FASTA sequences from NCBI
JGI >1000 metagenomes
•Filter
•Extract sequences
•Extract domains
Domain Positions
Domains FASTA
Domains Positions

Domains FASTA

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

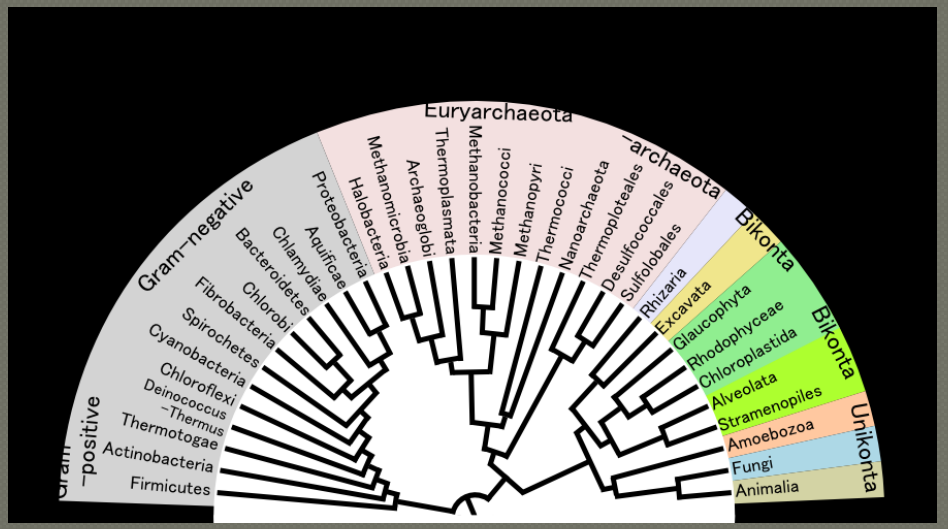
MAFFT

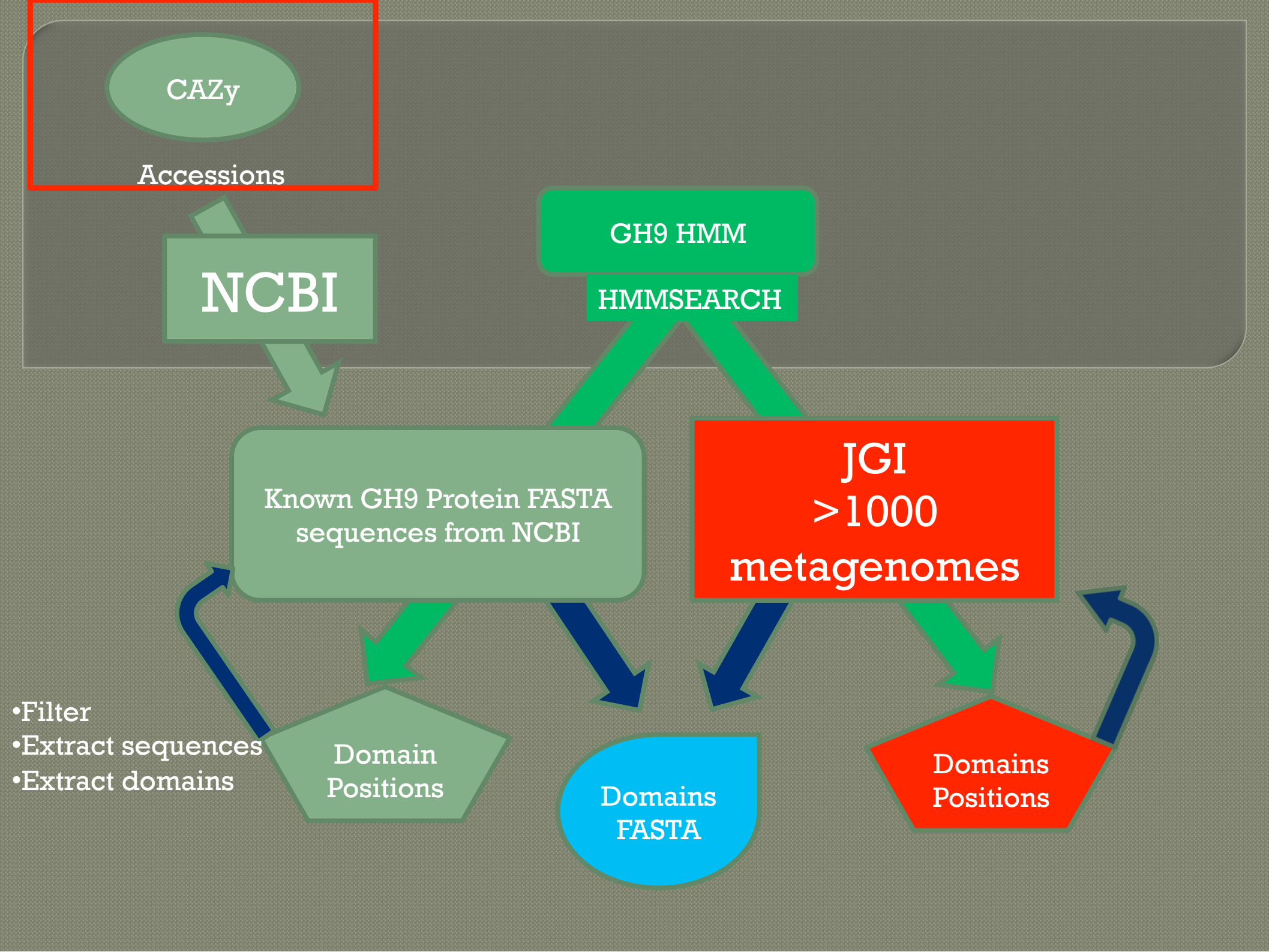
Domains Aligned

FastTree

ITOL

Phylogeny




CAZy
Accessions
NCBI
GH9 HMM
HMMSEARCH
Known GH9 Protein FASTA sequences from NCBI
JGI >1000 metagenomes
•Filter
•Extract sequences
•Extract domains
Domain Positions
Domains FASTA
Domains Positions

Dataset Download

Summary | All (1518) | Archaea (2) | Bacteria (530) | Eukaryota (779) | unclassified (207) | Structure (11) | Characterized (132)

< | 1 | 2 | >

Archaea

Protein Name	EC#	Organism	GenBank
Hmuk_0843		Halomicrobium mukohataei DSM 12286	ACV46974.1
Huta_2399		Halorhabdus utahensis DSM 12940	ACV12565.1

Bacteria

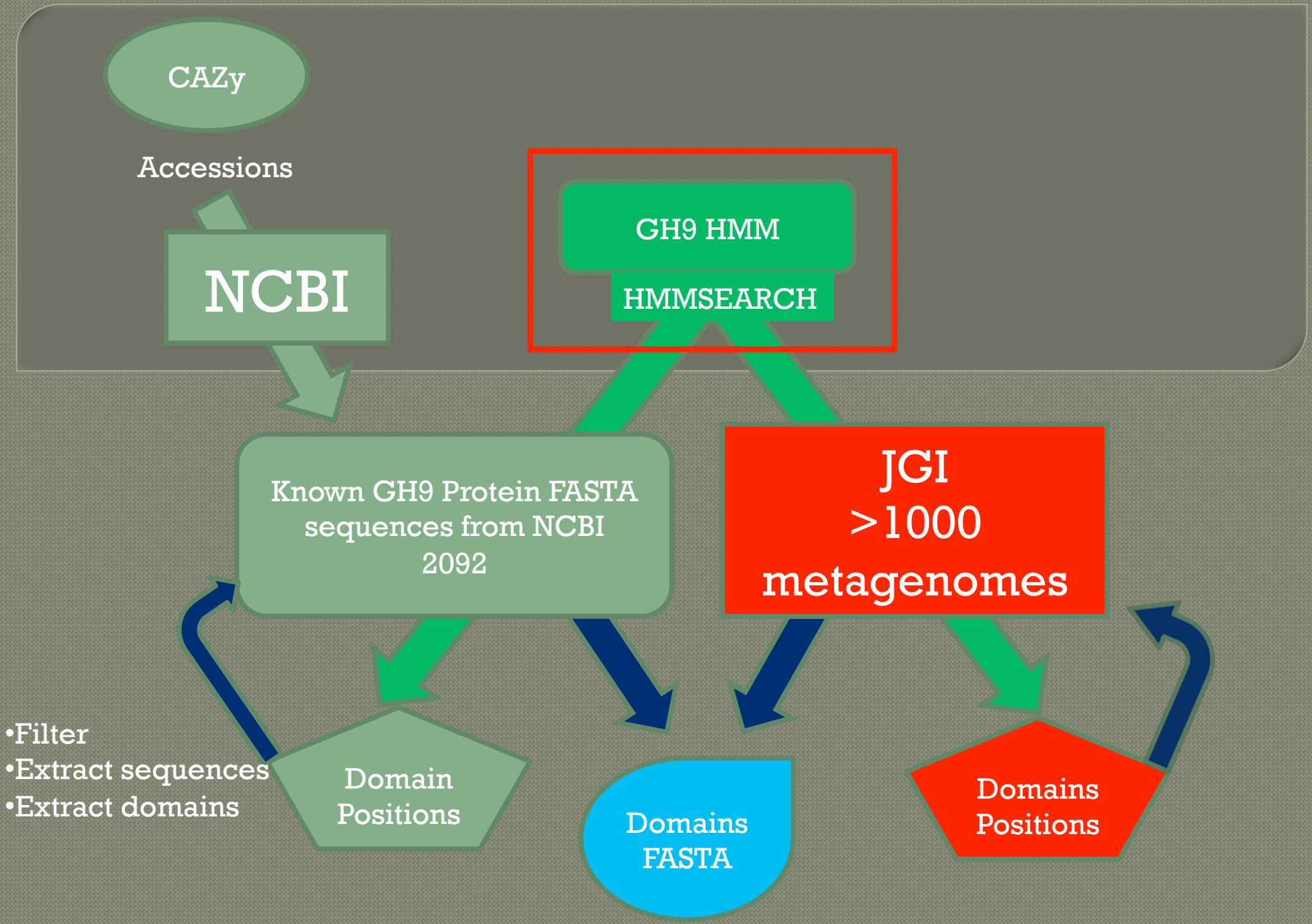
Protein Name	EC#	Organism	GenBank
AM1_2816		Acaryochloris marina MBIC11017	ABW27815.1
scaffoldin A (ScaA;CipV)		Acetivibrio cellulolyticus	AAF06064.1
endo-β-1,4-glucanase (Cel9B)	3.2.1.4	Acetivibrio cellulolyticus CD2 ATCC 33288	CAI94607.1
ACP_1014		Acidobacterium capsulatum ATCC 51196	
Acel_0970		Acidothermus cellulolyticus 11B ATCC 43068	ABK52743.1
Acel_1701		Acidothermus cellulolyticus 11B ATCC 43068	ABK53473.1
AMIS_26270		Actinoplanes missouriensis 431	BAL87847.1
AMIS_35140		Actinoplanes missouriensis 431	BAL88734.1
AMIS_48350		Actinoplanes missouriensis 431	BAL90055.1
ACPL_1206		Actinoplanes sp. SE50/110	AEV82103.1
ACPL_2890 (Cel1)		Actinoplanes sp. SE50/110	AEV83785.1
Amir_5339		Actinosynnema mirum DSM 43827	ACU39159.1
AHA_3098		Aeromonas hydrophila subsp. hydrophila ATCC 7966	ABK39743.1
β-glucosidase		Aeromonas hydrophila SUWA-9	BAF75999.1
ASA_3105		Aeromonas salmonicida subsp. salmonicida A449	ABO91099.1
B565_1044		Aeromonas veronii B565	AEB49079.1
cellulase (CelA;Aaci_2475) **(Cel9A)**	3.2.1.4	Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446	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

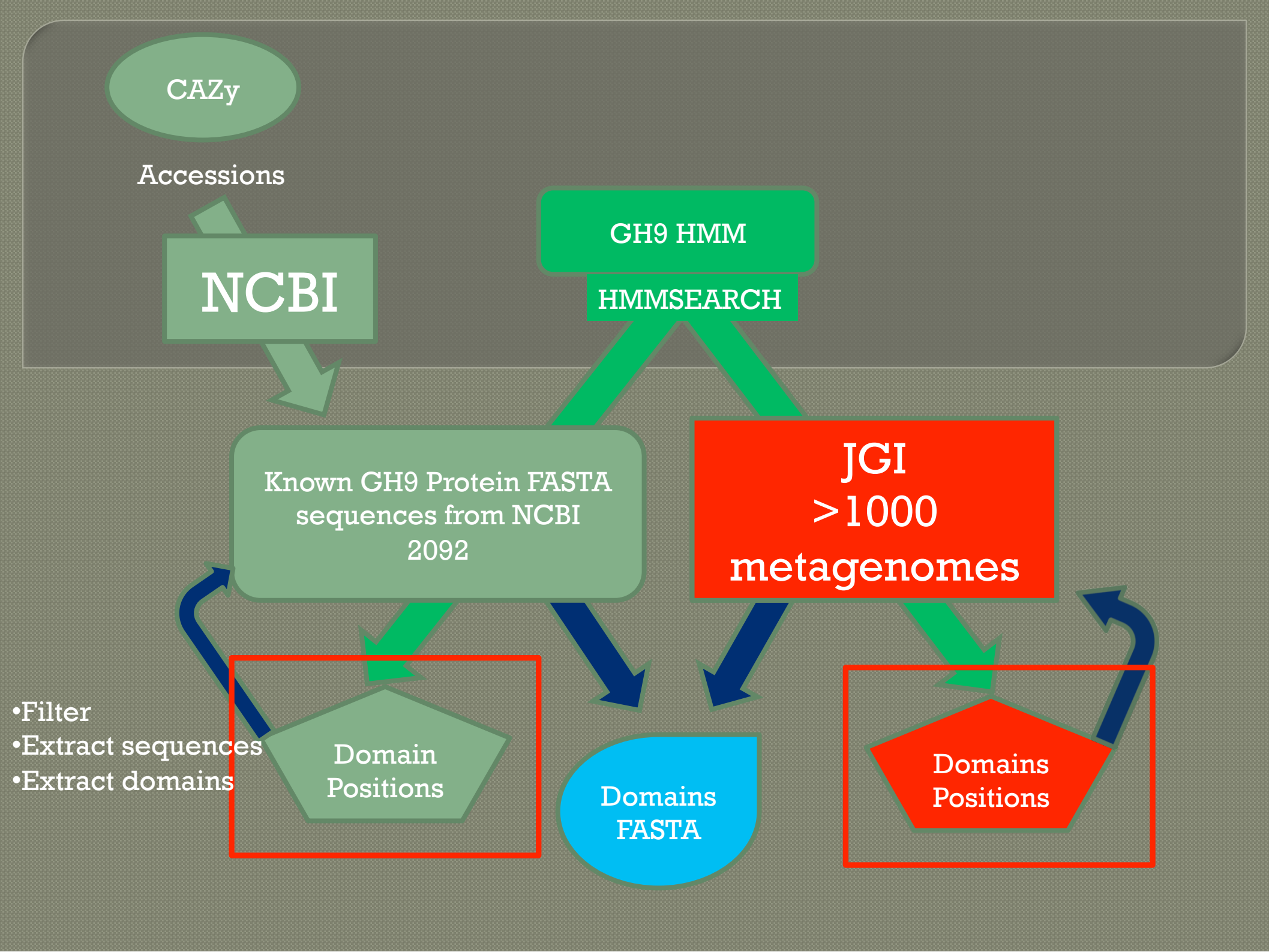
```
>gi|20415|emb|CAA39314.1| cellulase [Persea americana]
ASCGSTTVTAKNLISLAKKQVDYILGENPAKMSYMVGFGERYPQHVHHRGSSLPSVHAHPNPIPCNAGFQ
YLYSSSPNPNILVGAILGGPDSRDSFSDDRNNYQQSEPATYINAPLVGALAFFAANPVAN

>gi|20417|emb|CAA42569.1| cellulase [Persea americana]
MDCSSPLSLFHLLLVCTVMVKCCSASDLHYSDALEKSILFFEGQRSGKLPTNQRLTWRGDSGLSDGSSYH
VDLVGGYYDAGDNLKFGLPMAFTTTMLAWGIIEFGCLMPEQVENARAALRWSTDYLLKASTATSNSLYVQ
VGEPNADHRCWERPEDMDTPRNVYKVSTQNPGSDVAAETAAALAAASIVFGDSDSSYSTKLLHTAVKVFE
FADQYRGSYSDSLGSVVCPFYCSYSGYNDELLWGASWLHRASQNASYMTYIQSNGHTLGADDDDYSFSWD
DKRVGTKVLLSKGFLQDRIEELQLYKVHTDNYICSLIPGTSSFQAQYTPGGLLYKGSASNLQYVTSTAFL
LLTYANYLNSSGGHASCGTTTVTAKNLISLAKKQVDYILGQNPAKMSYMVGFGERYPQHVHHRGSSLPSV
QVHPNSIPCNAGFQYLYSSPPNPNILVGAILGGPDNRDSFSDDRNNYQQSEPATYINAPLVGALAFFAAN
PVTE

>gi|40672|emb|CAA28255.1| unnamed protein product [Clostridium thermocellum]
MSRMTLKSSMKKRVLSLLIAVVFLSLTGVFPSGLIETKVSAAKITENYQFDSRIRLNSIGFIPNHSKKAT
IAANCSTFYVVKEDGTIVYTGTATSMFDNDTKETVYIADFSSVNEEGTYYLAVPGVGKSVNFKIAMNVYE
DAFKTAMLGMYLLRCGTSVSATYNGIHYSHGPCHTNDAYLDYINGQHTKKDSTKGWHDAGDYNKYVVNAG
ITVGSMFLAWEHFKDQLEPVALEIPEKNNSIPDFLDELKYEIDWILTMQYPDGSGRVAHKVSTRNFGGFI
MPENEHDERFFVPWSSAATADFVAMTAMAARIFRPYDPQYAEKCINAAKVSYEFLKNNPANVFANQSGFS
TGEYATVSDADDRLWAAAEMWETLGDEEYLRDFENRAAQFSKKIEADFDWDNVANLGMFTYLLSERPGKN
PALVQSIKDSLLSTADSIVRTSQNHGYGRTLGTTYYWGCNGTVVRQTMILQVANKISPNNDYVNAALDAI
SHVFGRNYYNRSYVTGLGINPPMNPHDRRSGADGIWEPWPGYLVGGGWPGPKDWVDIQDSYQTNEIAINW
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
NCBI
GH9 HMM
HMMSEARCH
Known GH9 Protein FASTA sequences from NCBI 2092
JGI >1000 metagenomes
Domain Positions
Domains FASTA
Domains Positions
•Filter
•Extract sequences
•Extract domains

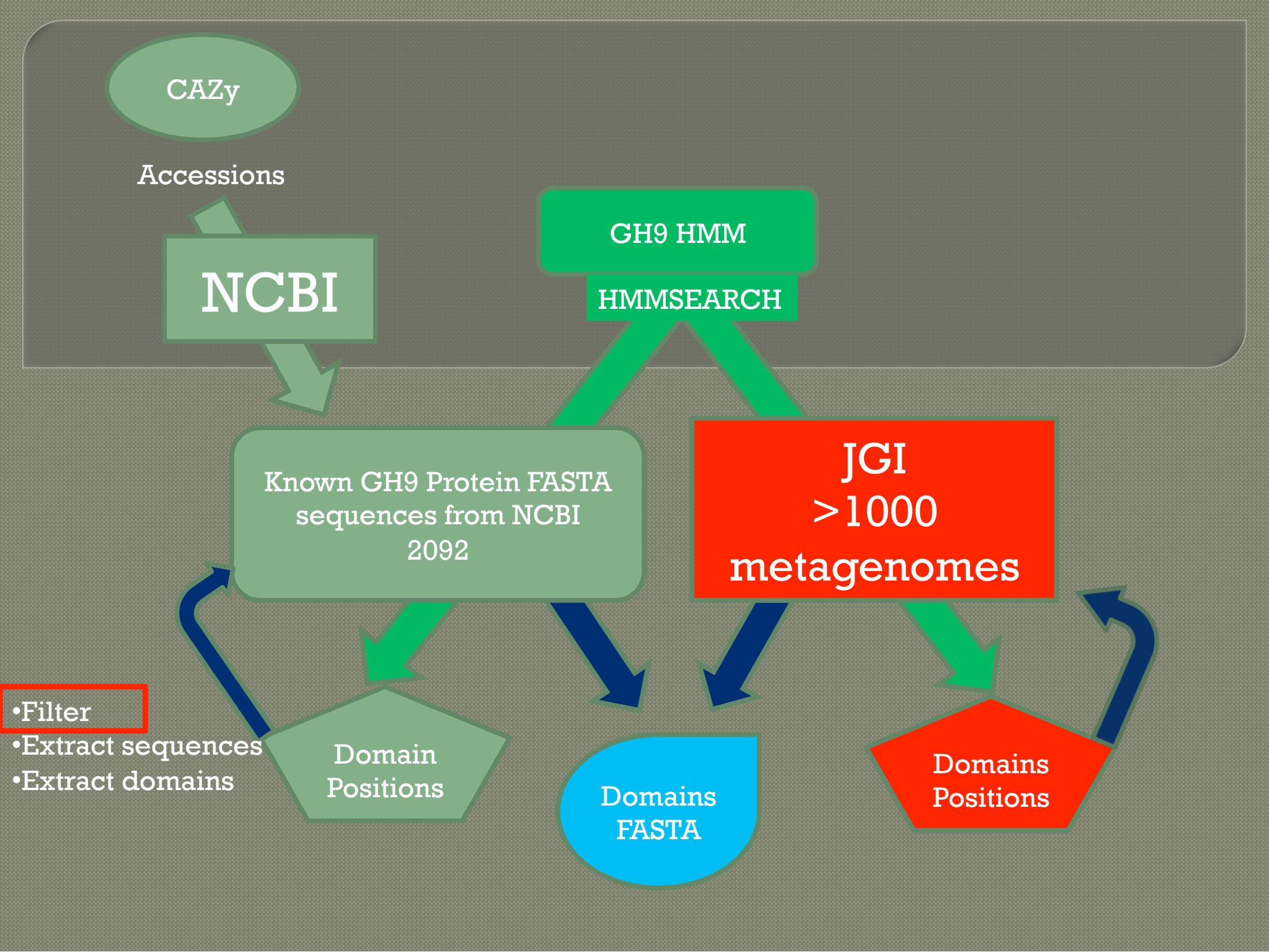
GH9 NCBI hmm – 2,102 Hits

```
                                                                                 --- full sequence --- ------------
omain -------------   hmm coord   ali coord   env coord
 target name                          accession   tlen query name           accession   qlen   E-value  score  bias   #  of  c-Ev
value  score  bias  from    to  from    to  from    to  acc description of target
                 ------------------- ---------- ----- -------------------- ---------- ----- --------- ------ ----- --- --- -----
-------- ------ ----- ----- ----- ----- ----- ----- ----- ---- ---------------------
i|42601450|gb|AAS21473.1|               -          1000 GH9.hmm              -            418  4.7e-193  639.2  21.3   1   2  3.9e
.9e-107  356.2   6.5     1   417    53   479    53   480 0.91 beta-1,4-endoglucanase 1 [Oikopleura dioica]
i|42601450|gb|AAS21473.1|               -          1000 GH9.hmm              -            418  4.7e-193  639.2  21.3   2   2   8.2
.2e-87  289.2   2.8     1   417   533   963   533   964 0.90 beta-1,4-endoglucanase 1 [Oikopleura dioica]
i|371940140|dbj|BAL45507.1|             -           653 GH9.hmm              -            418  4.7e-152  504.1  26.0   1   1  5.7e
.7e-152  503.9  18.0     1   418    41   475    41   475 0.94 glycoside hydrolase [Bacillus licheniformis]
i|264670906|gb|ACY72380.1|              -           646 GH9.hmm              -            418  1.1e-150  499.6  21.1   1   1  1.6e
.6e-150  499.1  14.6     1   418    34   467    34   467 0.93 cellulose hydrolase [Bacillus licheniformis]
i|39636954|gb|AAR29083.1|               -           633 GH9.hmm              -            418  2.4e-150  498.5  21.5   1   1  3.1e
.1e-150  498.2  14.9     1   418    21   454    21   454 0.93 cellulase [Bacillus licheniformis]
i|52003473|gb|AAU23415.1|               -           654 GH9.hmm              -            418  2.7e-150  498.3  21.5   1   1  3.4e
.4e-150  498.0  14.9     1   418    42   475    42   475 0.93 Glycoside Hydrolase Family 9 [Bacillus licheniformis DSM 13 = ATCC
i|52348141|gb|AAU40775.1|               -           654 GH9.hmm              -            418  2.7e-150  498.3  21.5   1   1  3.4e
.4e-150  498.0  14.9     1   418    42   475    42   475 0.93 cellulase EglA [Bacillus licheniformis DSM 13 = ATCC 14580]
i|4490766|emb|CAB38941.1|               -           997 GH9.hmm              -            418  1.5e-149  495.9  27.1   1   1  2.9e
.9e-149  494.9  18.8     1   418    40   479    40   479 0.95 cellulase [Paenibacillus barcinonensis]
i|37498962|gb|AAQ91573.1|               -           659 GH9.hmm              -            418  3.7e-149  494.6  19.9   1   1  4.7e
.7e-149  494.2  13.8     1   418    48   480    48   480 0.94 endoglucanase A precursor [Bacillus pumilus]
i|374343072|dbj|BAL46914.1|             -           997 GH9.hmm              -            418  4.8e-149  494.2  25.2   1   1  9.2e
.2e-149  493.3  17.5     1   418    40   479    40   479 0.95 beta-glucan hydrolase [Paenibacillus sp. WRL301]
i|188011009|gb|ACD44896.1|              -           616 GH9.hmm              -            418  1.6e-148  492.5  19.8   1   1  1.9e
.9e-148  492.2  13.7     1   418     5   437     5   437 0.94 endoglucanase [Bacillus pumilus]
i|310751795|gb|ADP09350.1|              -           659 GH9.hmm              -            418  1.6e-148  492.5  19.2   1   1  1.9e
.9e-148  492.2  13.3     1   418    48   480    48   480 0.94 endo-1,4-beta-glucanase [Bacillus pumilus]
i|148767913|gb|ABR10904.1|              -           659 GH9.hmm              -            418  1.9e-148  492.2  19.8   1   1  2.4e
.4e-148  491.9  13.7     1   418    48   480    48   480 0.94 endoglucanase [Bacillus pumilus]
i|6525242|gb|AAF15367.1|                -           659 GH9.hmm              -            418  1.9e-148  492.2  19.8   1   1  2.4e
.4e-148  491.9  13.7     1   418    48   480    48   480 0.94 endoglucanase [Bacillus pumilus]
i|122937813|gb|ABM68635.1|              -           659 GH9.hmm              -            418  1.9e-148  492.2  19.8   1   1  2.4e
.4e-148  491.9  13.7     1   418    48   480    48   480 0.94 endoglucanase A [Bacillus sp. AC-1]
i|1817723|gb|AAB42155.1|                -           880 GH9.hmm              -            418  4.9e-148  490.9  19.6   1   1  6.2e
.2e-148  490.6  13.6     1   418    51   484    51   484 0.94 beta-1,4-endoglucanase precursor [Thermobifida fusca YX]
i|71916307|gb|AAZ56209.1|               -           880 GH9.hmm              -            418  4.9e-148  490.9  19.6   1   1  6.2e
.2e-148  490.6  13.6     1   418    51   484    51   484 0.94 endoglucanase. Glycosyl Hydrolase family 9 [Thermobifida fusca YX]
```

GH9 JGI hmm – 15,428 Hits

```
#                                                                            --- full sequence --- -------------- this domain -------------   hmm coord   ali coord
env coord
# target name        accession   tlen query name           accession   qlen   E-value  score  bias   #  of  c-Evalue  i-Evalue  score  bias  from    to  from    to
from    to  acc description of target
#------------------- ---------- ----- -------------------- ---------- ----- --------- ------ ----- --- --- --------- --------- ------ ----- ----- ----- ----- -----
----- ----- ---- ---------------------
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 Cellobiohydrolase A (1,4-beta-cellobiosidase 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
```

Plain Text ▾ Tab Width: 8 ▾ Ln 1, Col 1 INS


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,$ARGV[0]);
print OUT "Hit ID\t\t\t\te-value\t\tbit score\tquery start\t query end\thit start\thit end\n";
while(<IN>){
if($_=~/#/){
$_=~s/^$//; #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

GH9.jgi.hmm.out.4.1(1).1

it 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

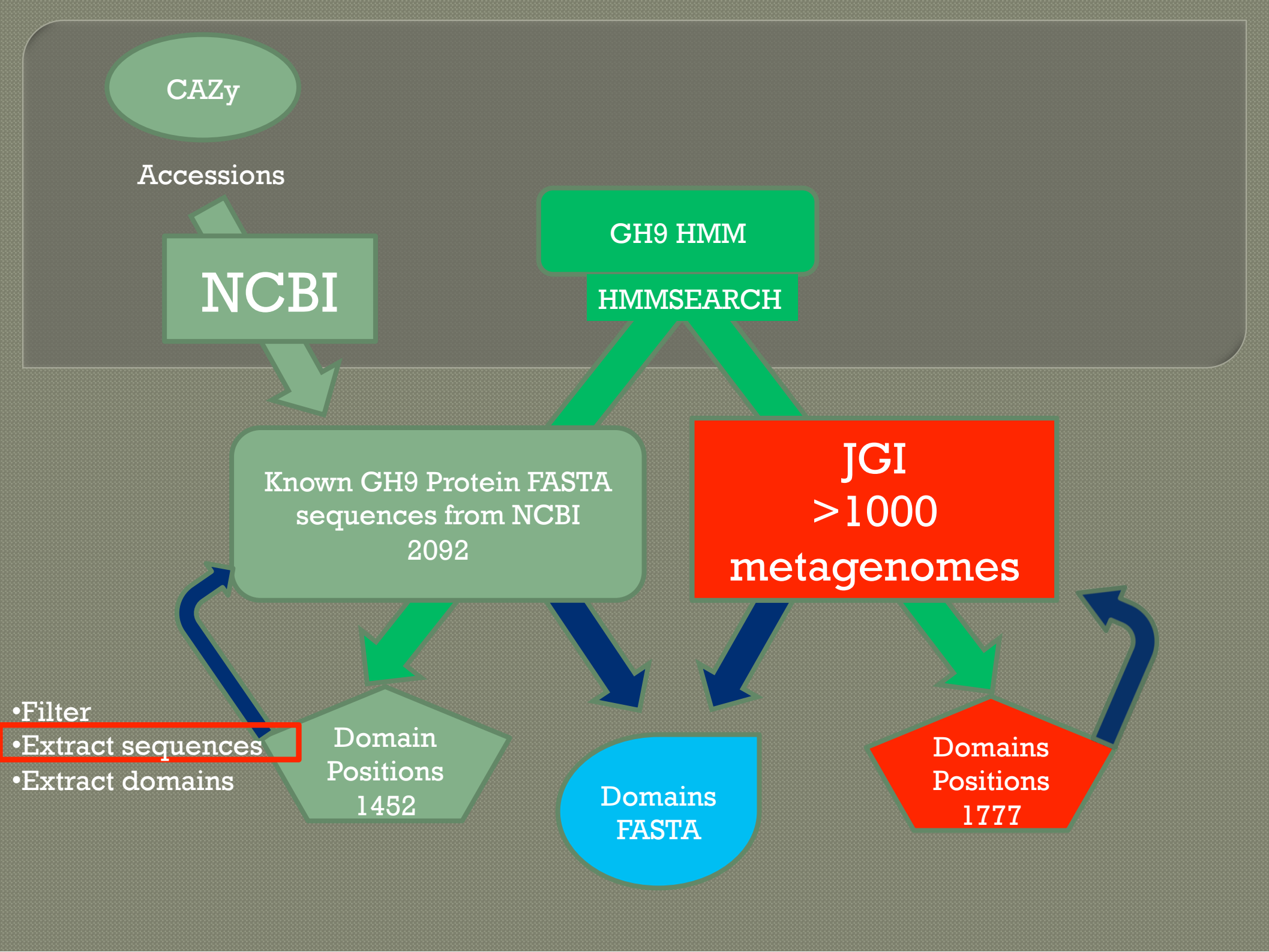
GH9 NCBI hmm out tabular format

it 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

Close document

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


CAZy
Accessions
NCBI
GH9 HMM
HMMSEARCH
Known GH9 Protein FASTA sequences from NCBI 2092
JGI >1000 metagenomes
•Filter
•Extract sequences
•Extract domains
Domain Positions 1452
Domains FASTA
Domains Positions 1777

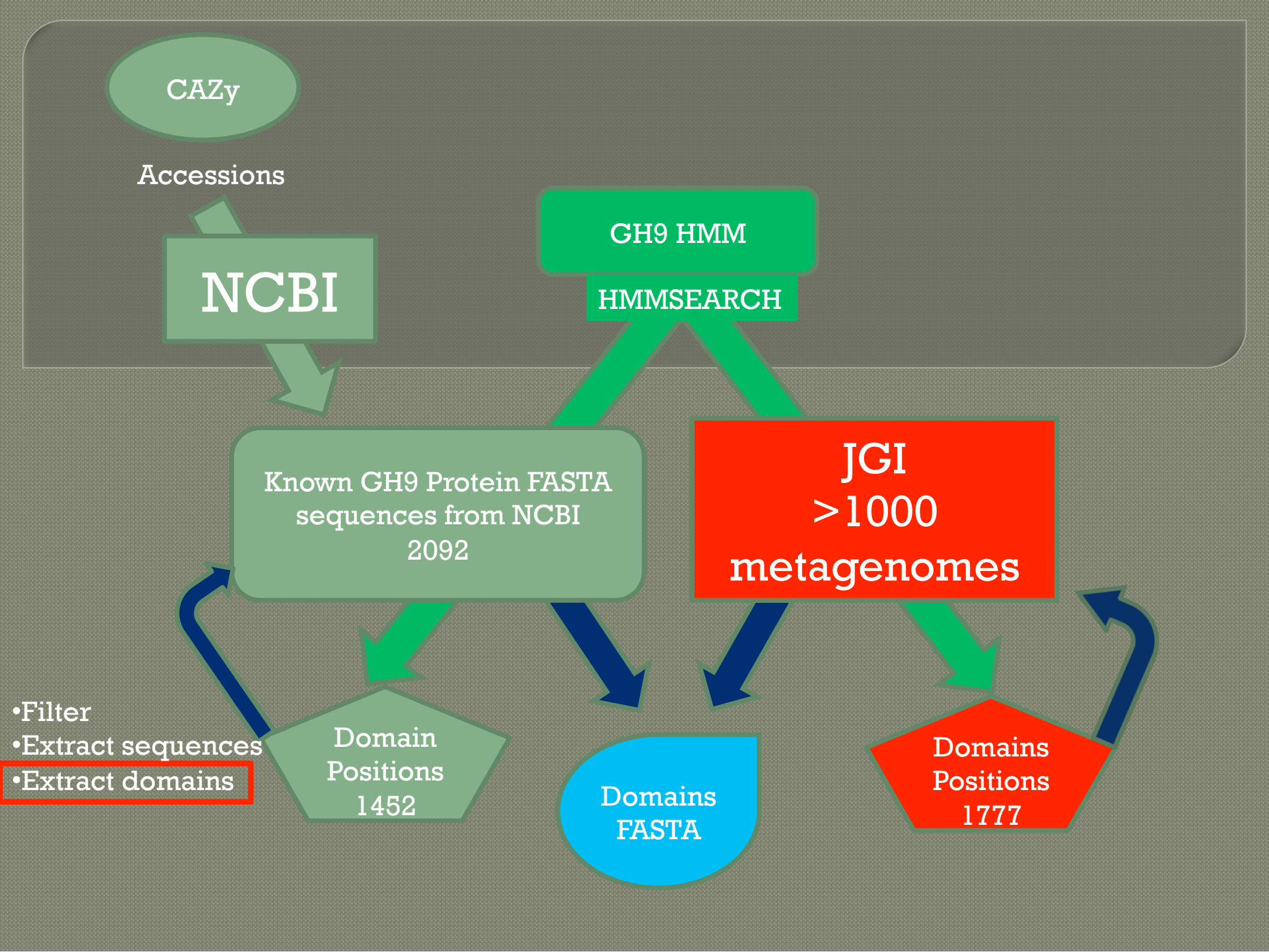
Perl script to extract full length protein FASTA sequences

get_sequence.pl retrieve_fasta.pl

```
#!/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;
```

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;
```

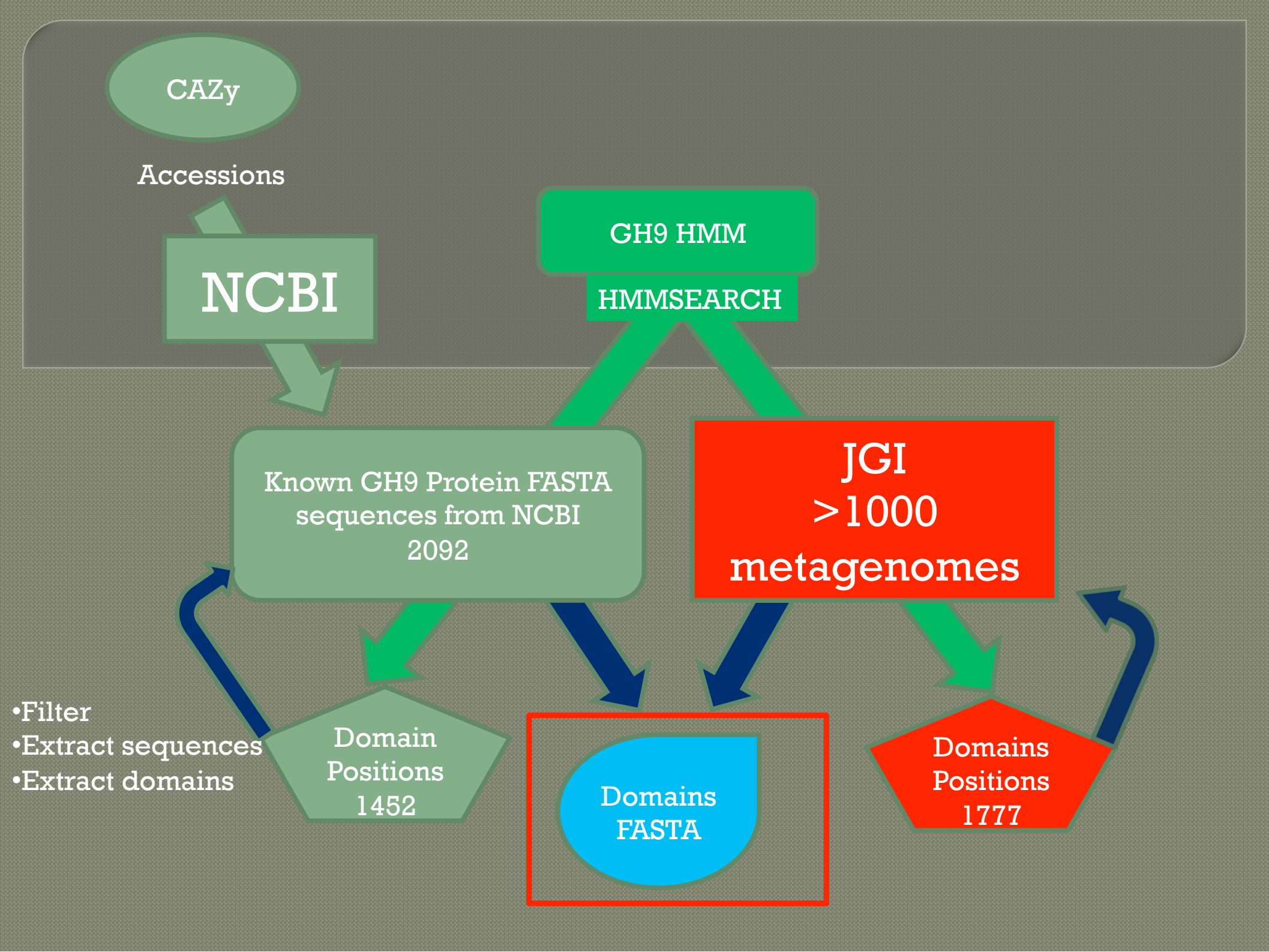

CAZy
Accessions
NCBI
GH9 HMM
HMMSEARCH
Known GH9 Protein FASTA sequences from NCBI 2092
JGI >1000 metagenomes
•Filter
•Extract sequences
•Extract domains
Domain Positions 1452
Domains FASTA
Domains Positions 1777

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:
```

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
NCBI
GH9 HMM
HMMSEARCH
Known GH9 Protein FASTA sequences from NCBI 2092
JGI >1000 metagenomes
•Filter
•Extract sequences
•Extract domains
Domain Positions 1452
Domains FASTA
Domains Positions 1777

All Domains in FASTA format

```
>gi|20417|emb|CAA42569.1||31|486|
DALEKSILFFEGQRSGKLPTNQRLTWRGDSGLSDGSSYHVDLVGGYYDAGDNLKFGLPMAFTTTMLAWGIIEE
>gi|40672|emb|CAA28255.1||140|570|
DAFKTAMLGMYLLRCGTSVSATYNGIHYSHGPCHTNDAYLDYINGQHTKKDSTKGWHDAGDYNKYVVNAGITV
>gi|45504|emb|CAA31082.1||126|596|
AAFYDALKYFYHNRSGIAIETPYTGGGRGSYASHSRWSRPAGHLNQGANKGDMNVPCWSGTCNYSLNVTKGWY
>gi|144416|gb|AAA23086.1||39|473|
AEALQKSMFFYQAQRSGDLPADFPVSWRGDSGLTDGADVGKDLTGGWYDAGDHVKFGFPMAFSATMLAWGAIE
>gi|144808|gb|AAA20892.1||78|510|
GEALQKAIFFYECQRSGKLDPSTLRLNWRGDSGLDDGKDAGIDLTGGWYDAGDHVKFNLPMSYSAAMLGWAVY
>gi|166947|gb|AAA32912.1||31|486|
DALEKSILFFEGQRSGKLPTNQRLTWRGDSGLSDGSSYHVDLVGGYYDAGDNLKFGLPMAFTTTMLAWGIIEE
>gi|167883|gb|AAA52077.1||27|452|
CSLLENALMFYKMNRAGRLPDNDIPWRGNSALNDASPNSAKDANGDGNLSGGYFDAGDGVKFGLPMAYSMTMI
>gi|296791|emb|CAA39010.1||32|462|
GEALQKAIMFYEFQRSGKLPENKRDNWRGDSGLNDGADVGLDLTGGWYDAGDHVKFNLPMAYSQTMLAWAAYE
>gi|305112|gb|AAA68129.1||103|551|
SNIKSFYYQRSGVELPERLAGIWARPAAHLDDKLEFHPTMERAGLWNAHGGWYDAGDYGKYIVNGGVSVATLM
>gi|310897|gb|AAC06387.1||278|746|
ELRVDALSFYYPQRSGIEILDSIAPGYGRPAGHIGVPPNQGDTDVPCAPGTCDYSLDVSGGWYDAGDHGKYVV
>gi|349601|gb|AAA02563.1||41|489|
ADALAKAILFFEGQRSGKLPSSQRVKWREDSALSDGKLQNVNLMGGYYDAGDNVKFGWPMAFSTSLLSWAAVE
>gi|530014|emb|CAA56918.1| 309 811
QMKYDALAFFYHKRSGIPIEMPYAGGEQWTRPAGHIGIEPNKGDTNVPTWPQDDEYAGIPQKNYTKDVTGGWY
```

Domains FASTA

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

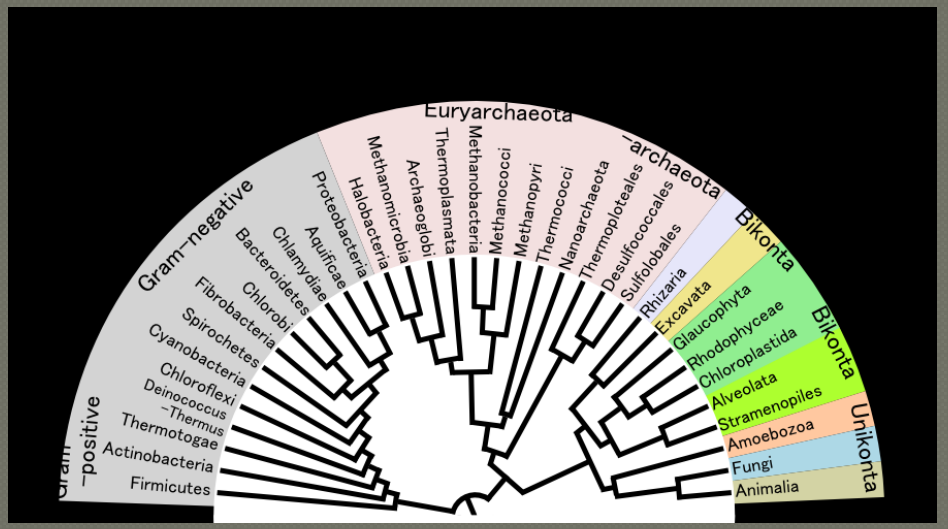
MAFFT

Domains Aligned

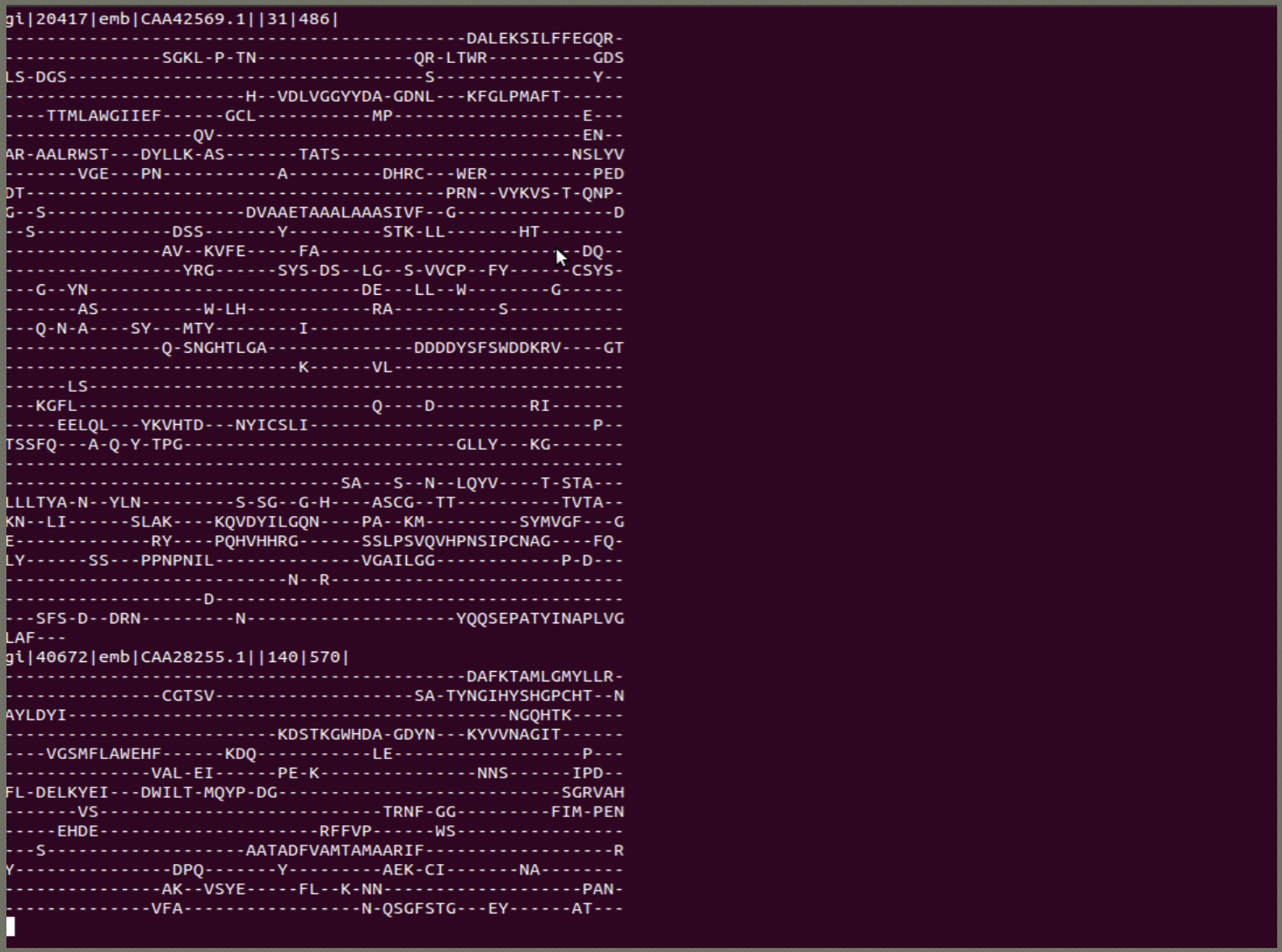
FastTree

ITOL

Phylogeny



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



Domains FASTA

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

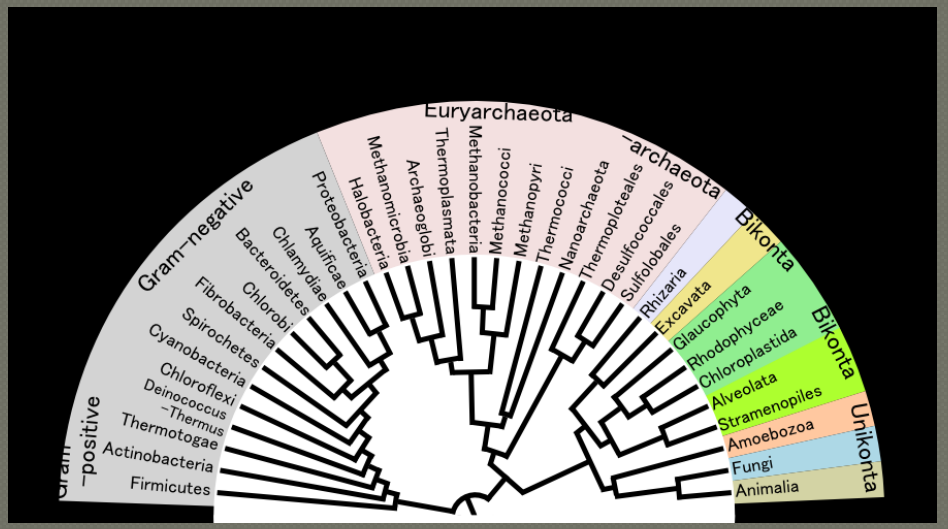
MAFFT

Domain s Aligned

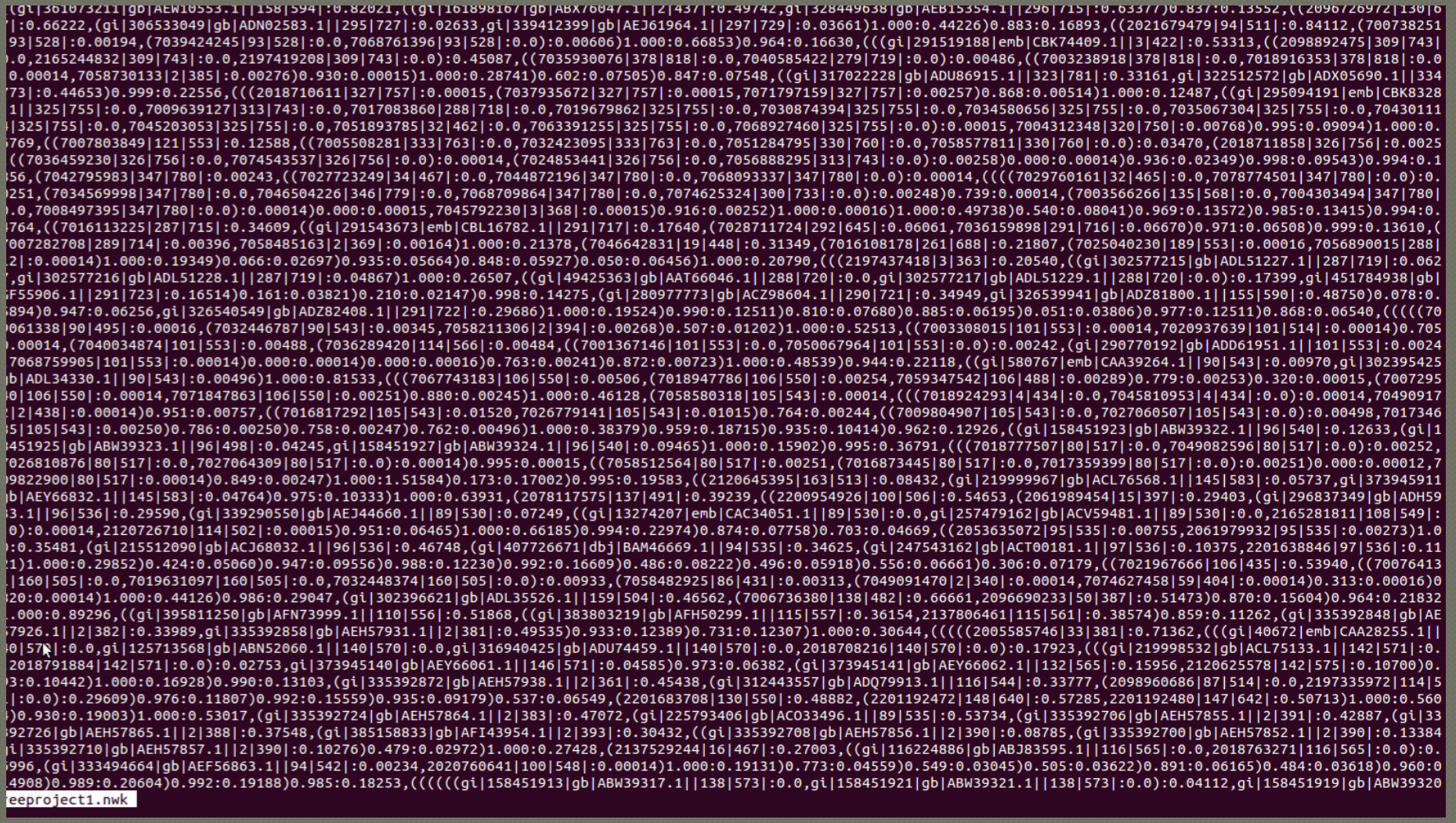
FastTree

ITOL

Phylogeny



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



Domains FASTA

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

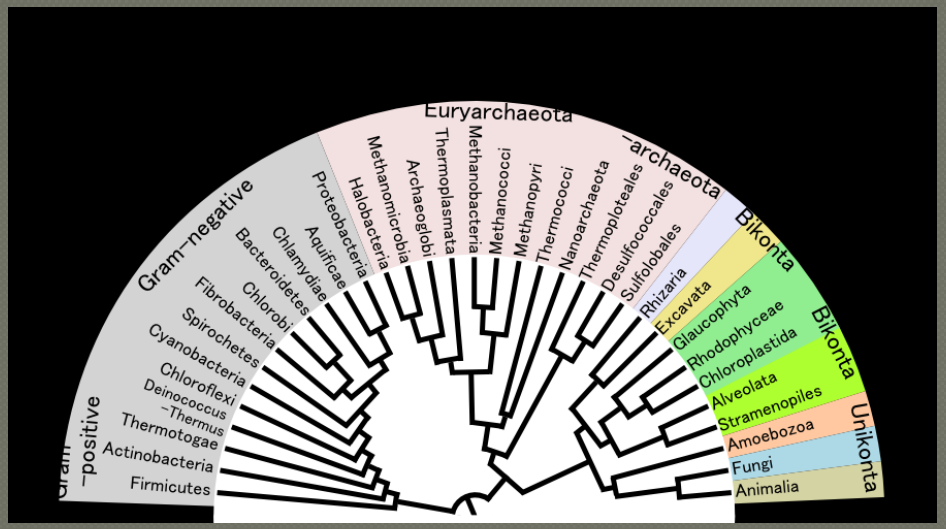
MAFFT

Domain s Aligned

FastTree

ITOL

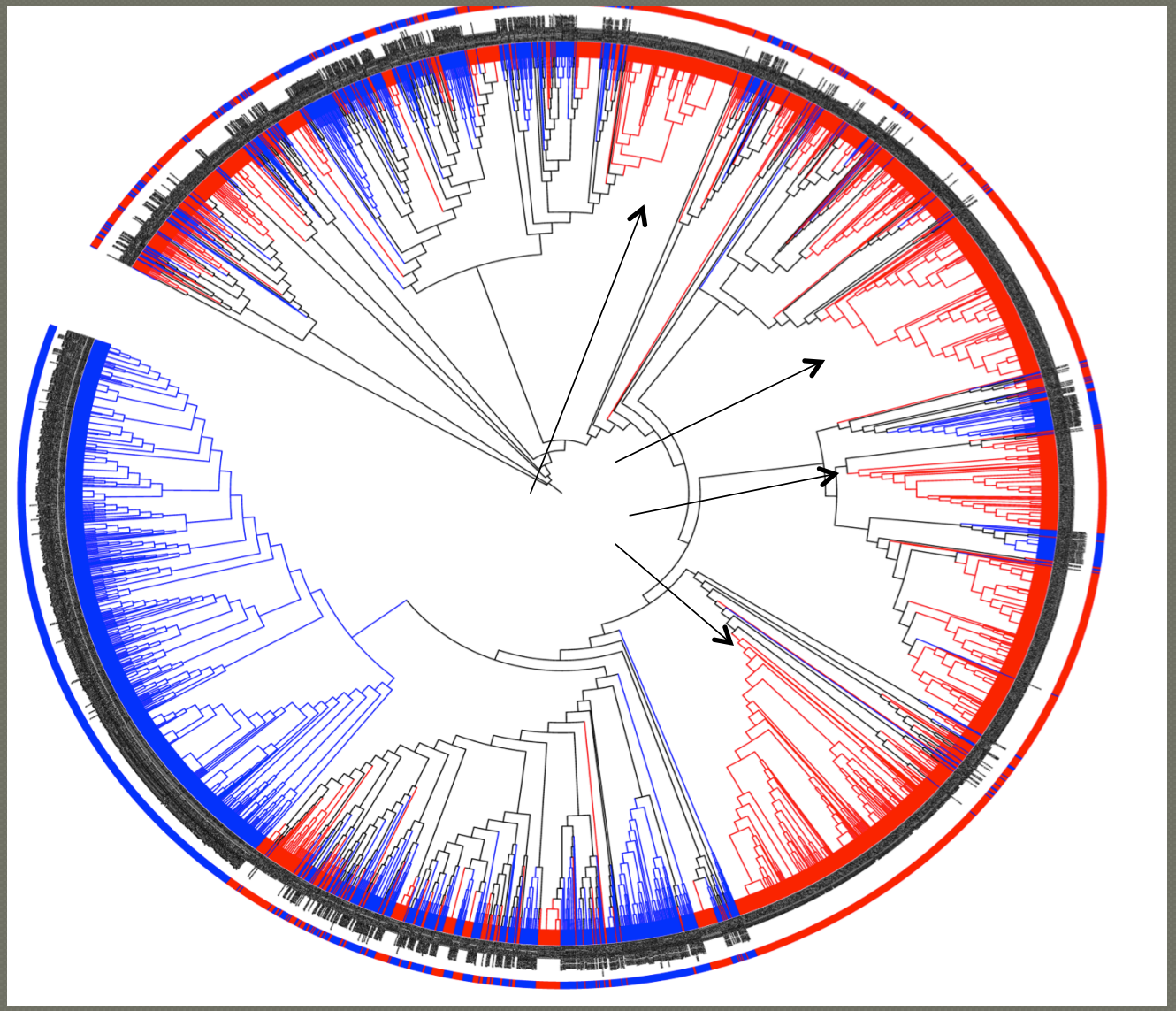
Phylogeny

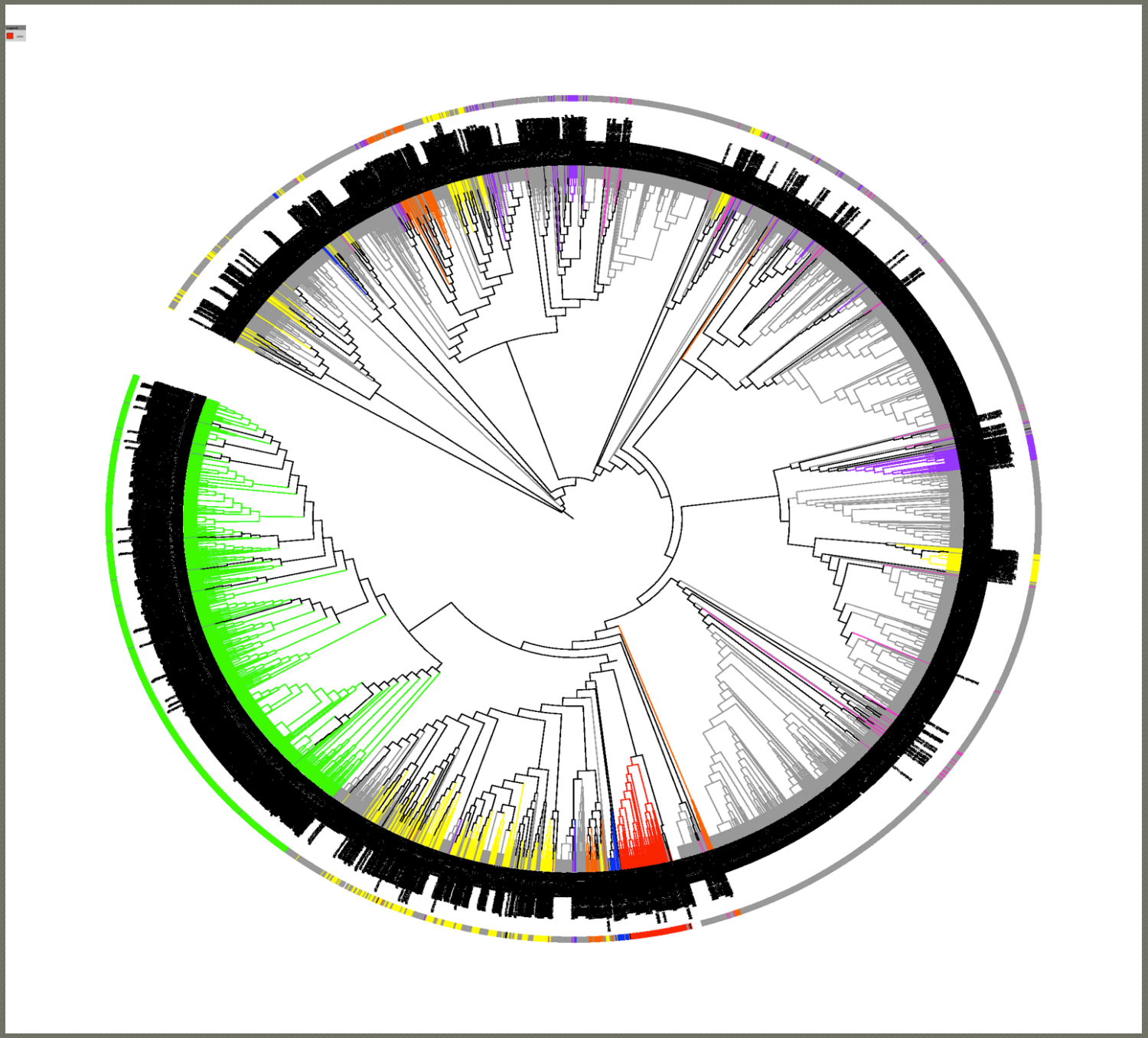


Tree created in iTOL-

Red: JGI

Blue: NCBI





Misc. Bacteria

Fungi

Cyanobacteria

Bacteroidetes

Metazoa

Actinobacteria

Proteobacteria

Viridiplante

Firmicutes

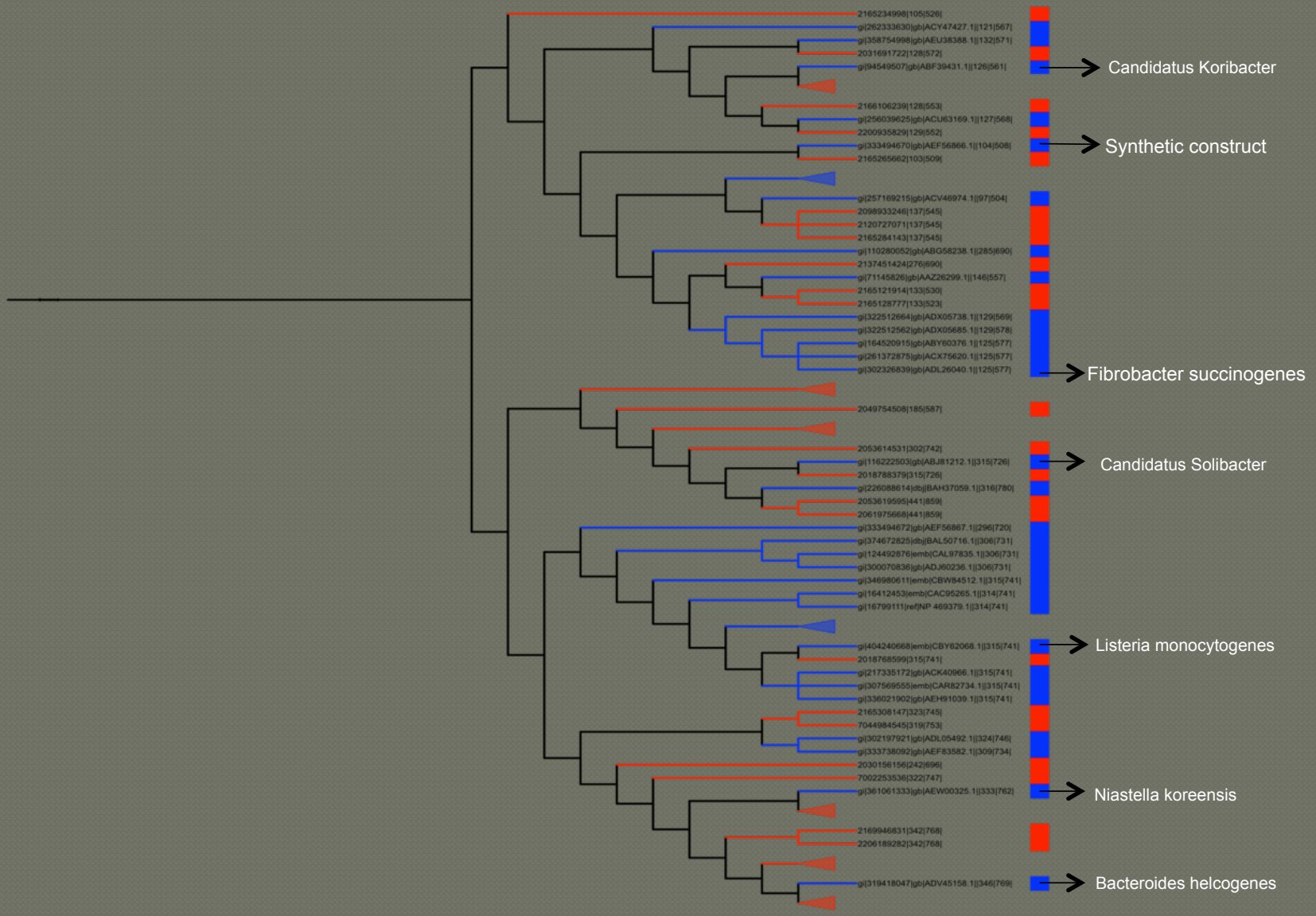
Protists

Stramenopiles

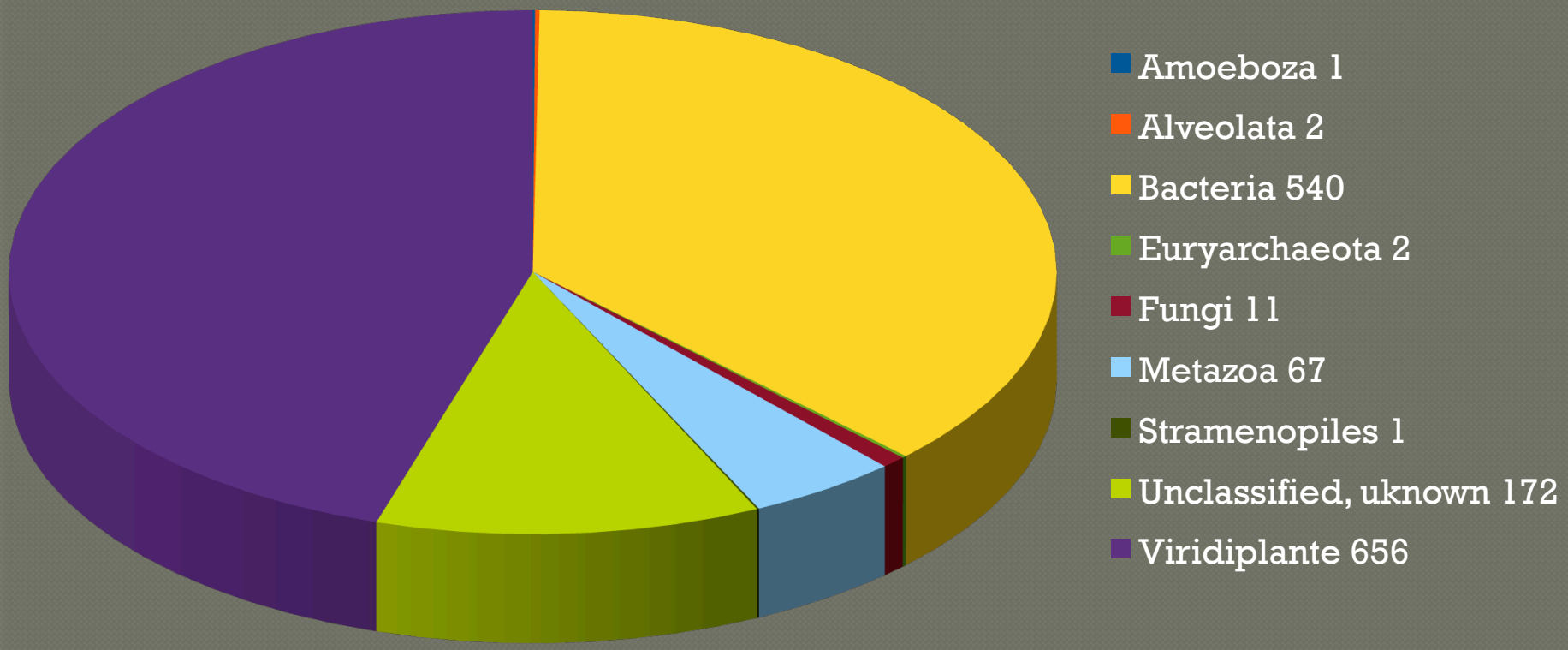
Euryarchaeota

All unknown organisms - Gray

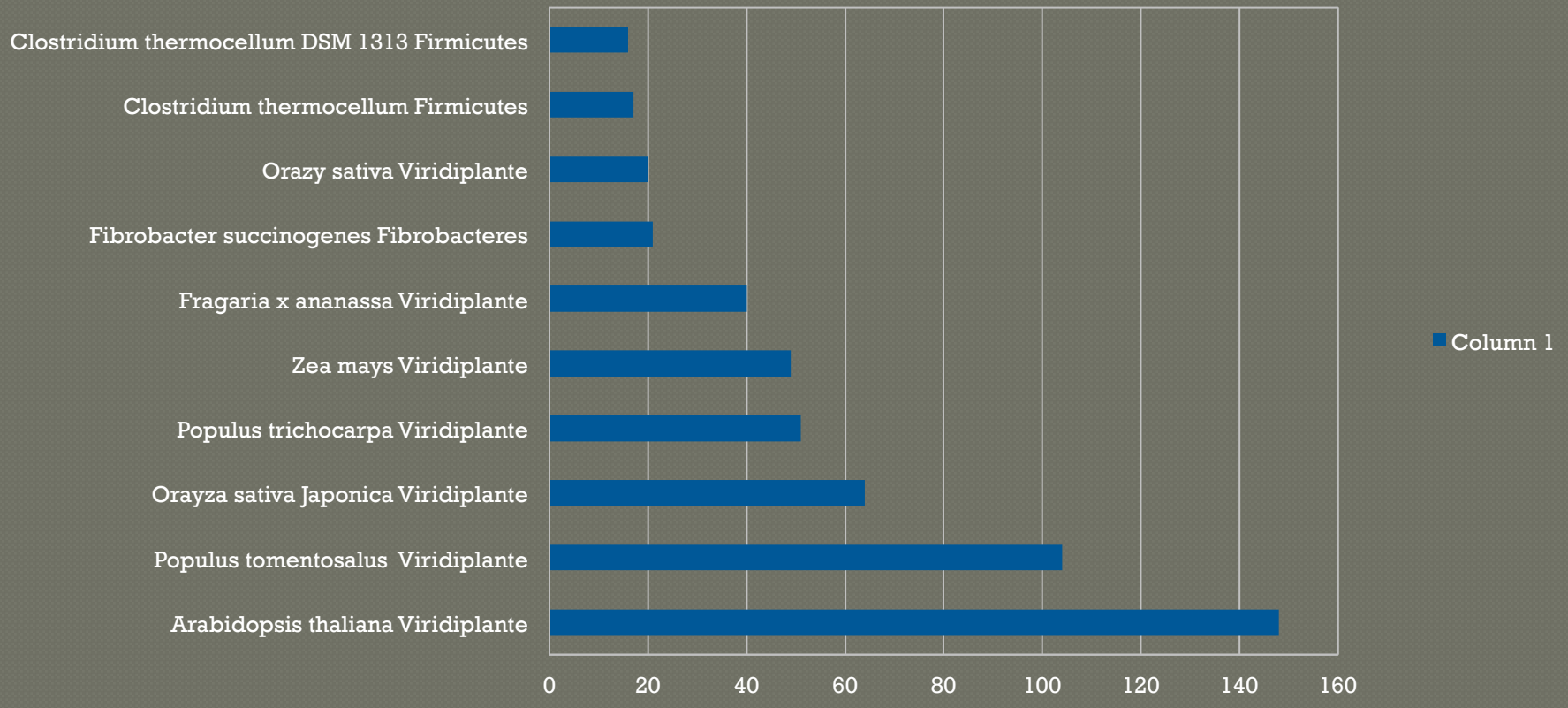
Un-banked GH9 proteins



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

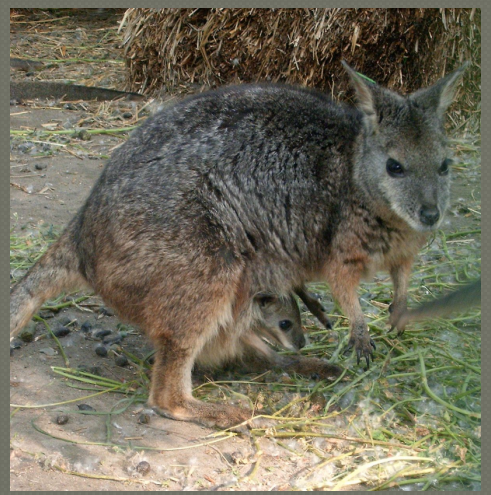


Top 10 organisms with multiple GH9 domains



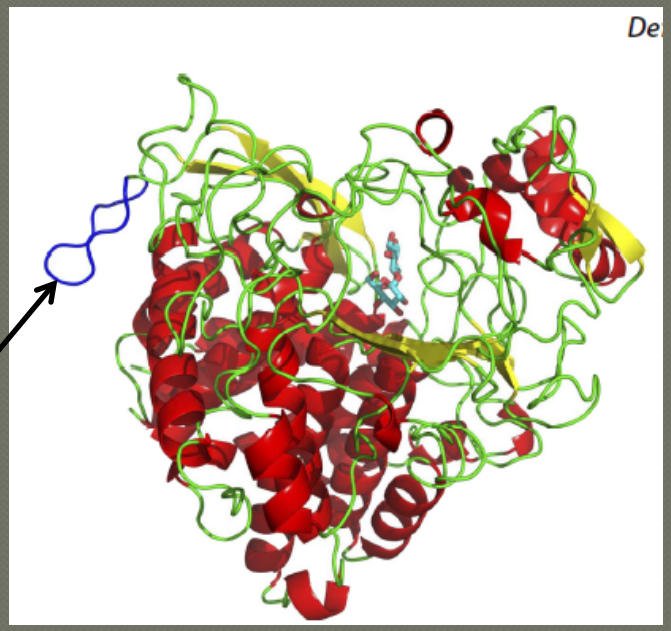
Examples of metagenomic sources of GH9 domains:

- 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.