Chelsea Murphy

Comparative Genomic Analysis of the Neocallimastigomycota CAZyome

ABSTRACT:

Members of the anaerobic fungi (Neocallimastigomycota) are known for their ability to degrade plant polysaccharides in the anoxic herbivorous gut. We aim to understand the evolutionary history, co-occurrence patterns and pangenomic diversity of the Carbohydrate Active Enzymes (CAZymes) repertoire in the Neocallimastigomycota. To this end, we sequenced the transcriptomes of fourteen different Neocallimastigomycota species. Those, along with four previously available transcriptomes, represent six different Neocallimastigomycota genera. From the transcriptomes, we extracted the protein families (pfam) domains encoding the glycoside hydrolases and polysaccharide lyases and analyzed the patterns of occurrence and phylogenetic diversity of these modules. The Neocallimastigomycota panCAZyome consisted of 128 families out of the 136 currently recognized, with the notable absence of the GH7 and GH61, both of which are ubiquitous in other fungal lineages. Interestingly, in spite of the large number of transcripts (465 to 1814) identified per transcriptome, it appears that all GH families consist of 1-5 distinct orthologues, suggesting massive gene duplication in the Neocallimastigomycota genomes. Further, within these orthologues, the majority appears to be obtained from bacterial donors by horizontal gene transfer, attesting to the importance of this process in conferring plant biomass degradation capacity and ecological fitness to this group of fungi. Most HGT events were observed in transcriptomes of all genera, suggesting the occurrence of this process prior to genus level diversification within this lineage. Major bacterial donors include lineages known to be prevalent in the herbivorous gut, e.g. Clostridiales,

Fibrobacteriales, and Ruminoccoccales, although additional events from lineages not associated with the herbivores guts were identified. Collectively, this work demonstrates the ancient origin and massive occurrence of HGT as a driving force behind the observed AGF efficient plant biomass degradation abilities.

INTRODUCTION:

The anaerobic gut fungi (AGF) belong to the deeply branched, highly distinct phylum Neocallimastigomycota. These microorganisms play an important role in the degradation of plant biomass in the anoxic herbivorous gut. Members of Neocallimastigomycota make up the only strictly anaerobic line of fungi (1). Due to the difficulty in culturing these organisms, not much research as has been performed on them to date. Our lab has particular experience with these unique organisms, as recent efforts by members of our lab resulted in the first published genome draft of an AGF isolate, *Pecoramyces ruminatum* sp. strain C1A (2).

For this project, our goal was to investigate instances of horizontal gene transfer (HGT) occurring within the different genera of the AGF. Through an analysis of the Carbohydrate Active Enzymes (CAZymes), multiple Glycoside Hydrolase (GH) genes were identified as having an apparent bacterial origin, suggesting they were acquired through the process of HGT. Analysis of the transcriptome suggested that these genes are actively transcribed, and play an important role in lignocellulosic biomass degradation in anaerobic fungi (3). The challenge of obtaining transcripts from this group for study has caused the extent of HGT in shaping lignocellulosic abilities across various AGF phyla to be currently unknown. In previous work, our lab has isolated a wide range of AGF belonging to all known genera, as well as some novel genera. We used this collection of genera to quantify HGT in AGF.

METHODS:

Phase I: Identifying GH families

Sixteen different isolates from six different genera were obtained, and had their RNA extracted with a yeast RNA extraction kit (Epicentre). These transcriptomes were then sequenced with Illumina MiSeq and assembled using Trinity. Gene calling was performed by Transdecoder. The glycoside hydrolase extraction was performed with an Hmm scan against the dbCan database.

Phase II: Horizontal Gene Transfer Analysis

Within each GH family detected, a Pfam extraction was performed. Sequences were clustered based on sequence similarity, and orthologues were obtained. Clustering of genes within the isolates was performed with an identity score of 0.90, and clusters formed from the supergroup across all genera were performed with an identity score of 0.99. Each of these orthologues then had their origin obtained through a Blastp against the nr database. Phylogenetic affiliation was then identified, by aligning to a reference pfam alignment and the construction of trees based on sequence similarity. Through these steps, the origins of the genes were able to be identified.

RESULTS:

Phase I: Identification of GH families

The number of transcripts for each isolate, both total and belonging to a GH family, are shown below in Table 1. The number of total number of transcripts varied from 14,648 in the PirE2 *Piromyces* isolate to 57,019 in the R3 Novel Genus isolate, with a mean score of 33,459 and median of 30,631. The numbers of GH transcripts ranged from 465 in the *Anaeromyces* isolate from the *Anaeromyces* genus to 1,814 in the R3 Novel Genus isolate, with a mean score of 1,006

and median of 1,052. The percentage of the transcriptome that the GH families made up ranged from 2.26% in the C3G *Anaeromyces* isolate to 4.83% in the PirE2 *Piromyces* isolate, with a mean score of 3.15% and median of 3.21%.

Isolate name	Genus	Source	Total # of transcripts	# of transcripts with GH family domain
C1A	*Pecoramyces*	OSU (2)	16347	531
S4B		OSU	40003	1120
C3G	*Anaeromyces*	OSU	28392	643
C3J		OSU	51884	1238
O2		OSU	56039	1472
Na		OSU	44051	1423
Anaeromyces		CA (4)	15285	465
R3	Novel genus	OSU	57019	1814
D3A	*Orpinomyces*	OSU	31661	762
D4C		OSU	29660	1007
D3B		OSU	29475	1237
G3	*Neocallimastix*	OSU	46303	1101
Neocallimastix		CA (4)	23933	849
R2	*Piromyces*	OSU	31602	1096
Piromyces		CA (4)	19047	633
PirE2		JGI	14648	707

Table 1. Numbers of transcripts.

The distribution of the GH transcripts across each family can be seen in Figure 1. There were some universal trends in the number of the transcribed GHs across all the genera, such as GH5. Typically, the numbers of GHs transcribed in each family exhibited greater similarity in peaks within the genera, with discrepancies in the relative sizes of the peaks due to differences in the overall number of transcripts obtained from each isolate. Note the unusual lack of GH7 and GH61, families ubiquitous in other fungi.

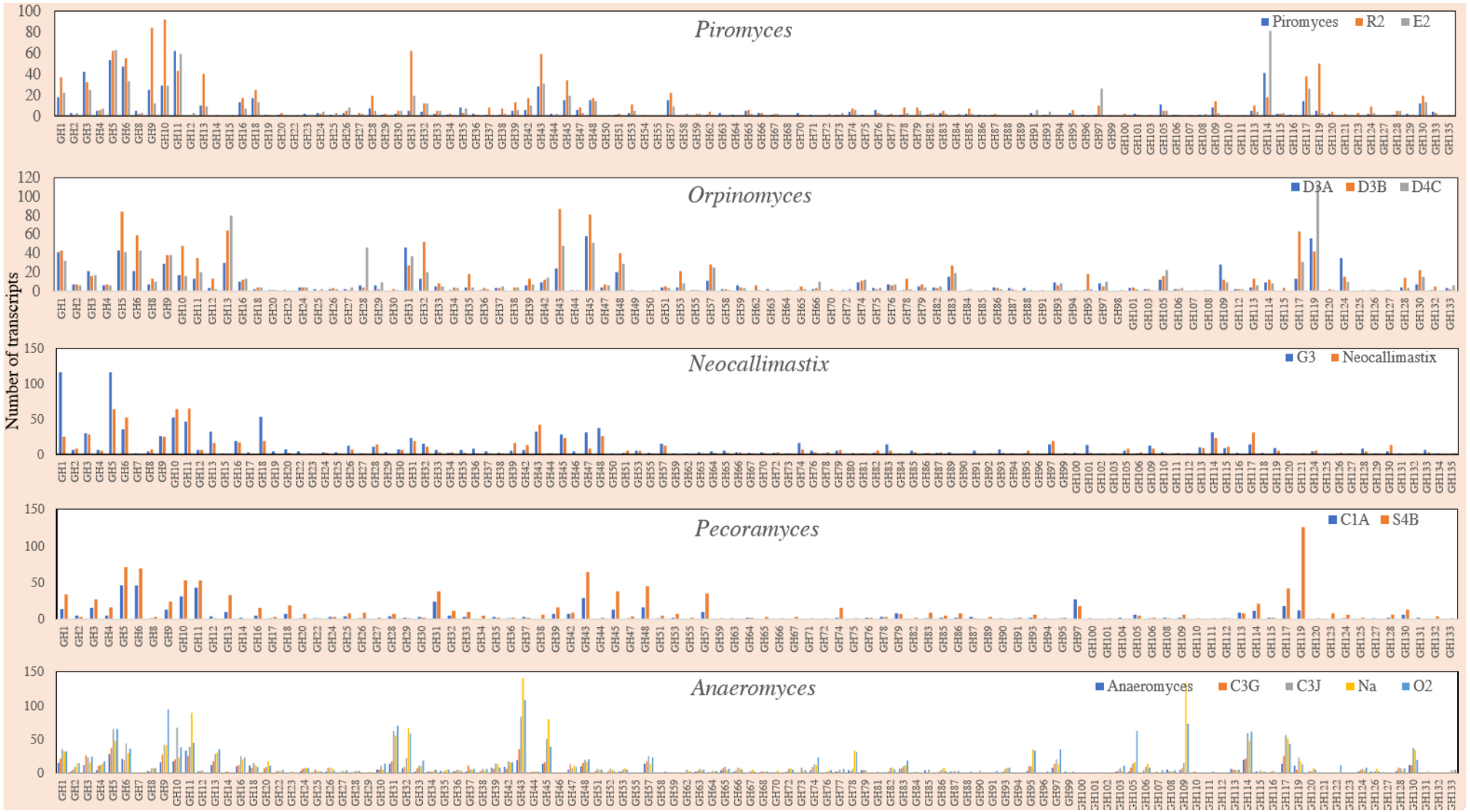

Figure 1. Distribution of transcripts.

Phase II: Horizontal Gene Transfer Analysis

The results of the preliminary analysis of HGT events for various GH families can be seen in Table 2. Of the 44 families analyzed, 36 (81.8%) show evidence of HGT. Families that have been analyzed so far were selected because at each isolate had at least one transcript from that GH family, allowing for the general trend of HGT to be more easily seen and supported. The bacterial origins come from a variety of different taxa, ranging from the Family to Phyla level.

GH family	Evidence of HGT	HGT donors
GH1	Yes (partial)	Family Clostridiaceaea
GH2	Yes (All)	Family Clostridiaceaea
GH3	Yes (partial)	Family Ruminococciaceae
GH5	Yes (All)	Order Clostridiales
GH6	Yes (partial)	Class delta Proteobacteria
GH8	Yes (All)	Order Fibrobacterales
GH9	Yes (Partial)	Order Actinomycetales
GH10	Yes (Partial)	Oder Thermales
GH11	Yes (All)	Order Fibrobacterales
GH13	Yes (partial)	Order Clostridiales, Order Lactobacillales
GH16	Yes (Partial)	Order Clostridiales
GH18	Yes (Partial)	Order Bacillales
GH20	No	
GH24	Yes (Partial)	Phylum Proteobacteria
GH26	Yes (All)	Order Fibrobacterales
GH28	Yes (partial)	Order Actinomycetales
GH31	No	
GH32	Yes (All)	Order Clostridiales
GH33	Yes (partial)	Order Actinomycetales
GH35	Yes (All)	Order Clostridiales
GH37	No	
GH39	Yes (All)	Order Clostridiales
GH42	Yes (All)	Family Ruminococciaceae, Order Actinomycetales
GH43	Yes (All)	Order Clostridiales, Order Bacteroidales
GH45	No	
GH47	No	
GH48	Yes (All)	Order Clostridiales, Order Actinomycetales
GH53	Yes (All)	Order Clostridiales
GH57	No	
GH74	Yes (Partial)	Family Ruminococciaceae
GH78	Yes (All)	Order Bacteroidales
GH87	No	
GH93	Yes (All)	Order Actinomycetales
GH95	Yes (All)	Order Clostridiales, Order Bacteroidales
GH97	No	
GH105	Yes (partial)	Order Actinomycetales
GH106	Yes (All)	Order Ruminococciaceae
GH109	Yes (partial)	Order Clostridiales
GH113	Yes (All)	Order Clostridiales, Order Thermales
GH117	Yes (All)	Order Actinomycetales
GH119	Yes (All)	Order Actinomycetales
GH124	Yes (All)	Order Deinoccales
GH128	Yes (All)	Order Clostridiales, Order Actinomycetales
GH130	Yes (All)	Order Clostridiales

Table 2. Preliminary HGT analysis of selected GH families.

DISCUSSION:

Out of all the different transcripts obtained from each organism, a large number of them belonged to the GH family. Due to their high importance in increasing the fitness of these fungi in the highly competitive environment of the anaerobic gut, this is to be expected. Not only were transcripts from the GH families obtained in high number, but in high variety as well. The Neocallimastigomycota panCAZyome consisted of 112 out of the 136 currently recognized GH families (SOURCE 5). Similarities in the frequency of occurrence of each GH family can be seen across all of the different isolates, with isolates within the same genera being even more closely related. As our data was collected from transcripts, we know that all of these GH families were being used by the organisms, but some appear to be more important, and thus more frequently transcribed, than others.

When clustering into orthologues is performed, an interesting trend emerged. While the number of GHs was vast, most all families clustered into 1-5 distinct orthologues. This suggests that the genes were duplicated on a massive scale post uptake, once again demonstrating their importance to survival in the anaerobic gut. In addition, a majority of orthologues appear to have a bacterial origin, suggesting acquisition of these genes via horizontal transfer from bacteria that were already thriving inside the anaerobic gut when the fungi were introduced. The bacterial donors for these genes appear to largely be taxa known to be found in the gut, such as Clostridiales, Fibrobacteriales, and Ruminoccccoccales. There were some HGT events that appeared to stem from other bacteria which are not indigenous to the herbivorous gut, such as Order Thermales, so further work will need to be done to verify these events and reconcile them with the evolutionary history of the AGF. Of the genes that did show evidence of HGT, most events occurred across all of the genera. This suggests that the various GH families were

obtained prior to genus level diversification within the phyla. This timing makes sense, as the GH families encode massively important genes that would need to be acquired right away for survival, long before mutations have a chance to cause significant divergence.

Further work is being performed to further validate and expound upon these results. While only 6 out of the 9 genera of Neocallimastigomycota were analyzed for this study, more transcriptomes are being sequenced and analyzed to not only allow for more data within each genus to be collected, but to allow for more genera to be analyzed. In addition, HGT events will be confirmed through phylogenetic placements of transcripts using alien index measurements, and transcripts will be blasted on several different databases to determine the organism classification that most closely matches the gene. Once work on the GH families is complete, the same methods will be employed to analyze other functionally important families, such as the polysaccharide lyases (PLs).

Collectively, the results of this study indicate the ancient origin of genes crucial to plant biomass degradation in the anaerobic herbivorous gut. HGT appears to have played a massive role in the ecological fitness of the AGF, as an extremely high number of GH families appear to have a bacterial origin. These critical genes were taken up shortly after introduction into the gut environment, prior to further diversification of the phyla, and have been duplicated in large scale since.

REFERENCES:

1. FEMS Microbiol. Ecol., 2014. **90**:1-17.
2. Appl. Environ. Microbiol., 2013. **79**:4620-4634.
3. Biotechnol. Biofuels, 2015. **8**:208.

4. Science, 2016. **351**:1192-1195.

5. Nucl. Acid. Res., 2014. **42**:D490-495.