MycoCosm: Comparative analysis of Gene Families

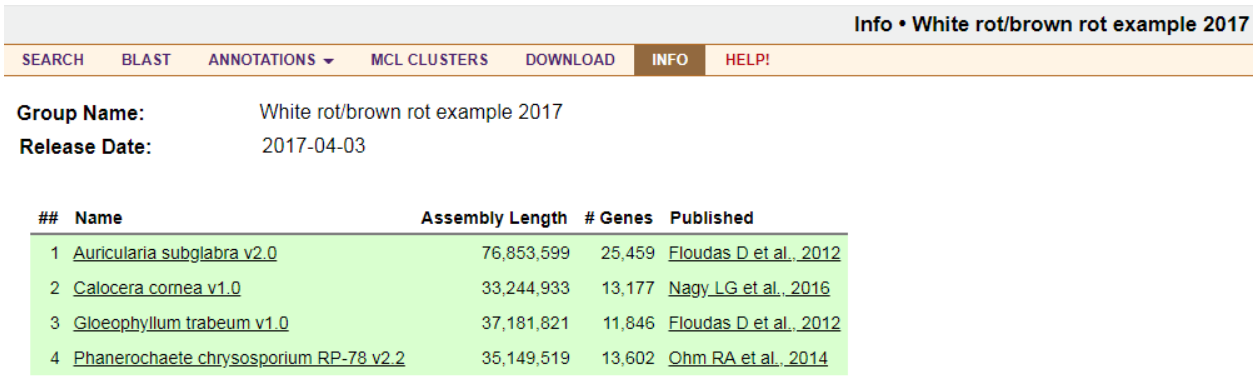
Objective: Compare genomes of wood decay fungi to identify gene families which can be used to distinguish white rot and brown rot fungi

Many fungi of the phylum Basidiomycota are capable of degrading wood, including the recalcitrant polymer lignin, which gives wood its structural strength and resistance to microbial attack (Floudas et al. 2012; Riley et al. 2014). These wood decaying fungi are often classified as either white rot, in which lignin is completely degraded and cellulose is left somewhat intact; or brown rot, in which cellulose is degraded and lignin left somewhat intact. While the precise enzymatic mechanisms vary from one fungus to another, in general the white rot fungi's genomes encode class II peroxidase enzymes (CAZy: AA2) to break down lignin; carbohydrate-binding motifs (CAZy: CBM1) to bind cellulose; and glycoside hydrolases of families 6 and 7 (CAZY: GH6 and GH7) to break down cellulose. The genome of a brown-rot fungus tend to lack genes encoding these enzymes, or have them in reduced numbers compared to white rot fungi.

Suppose we are comparing the genomes of four wood decaying fungi: *Auricularia subglabra*, *Calocera cornea*, *Gloeophyllum trabeum*, *Phanerochaete chrysosporium* RP-78. Suppose, also, that we don't know which of them are white-rot or brown-rot fungi. How can we use MycoCosm to make predictions about their mode of decay?

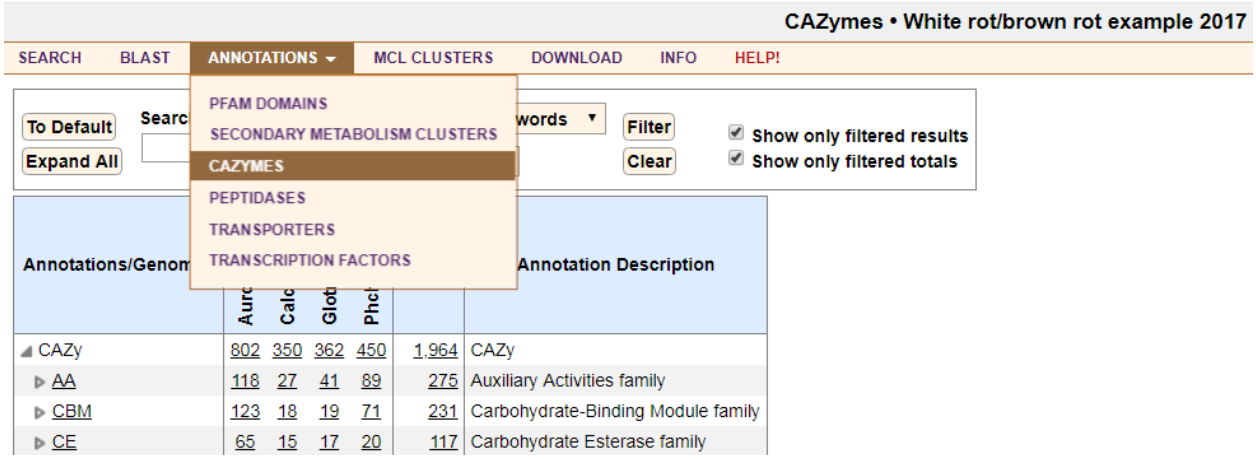
Start by going to the genome group page created for this example (in real life we would use a similar genome group page, but with a larger, ecologically- or phylogenetically-relevant selection of organisms):

https://genome.jgi.doe.gov/WR_BR_example_2017/

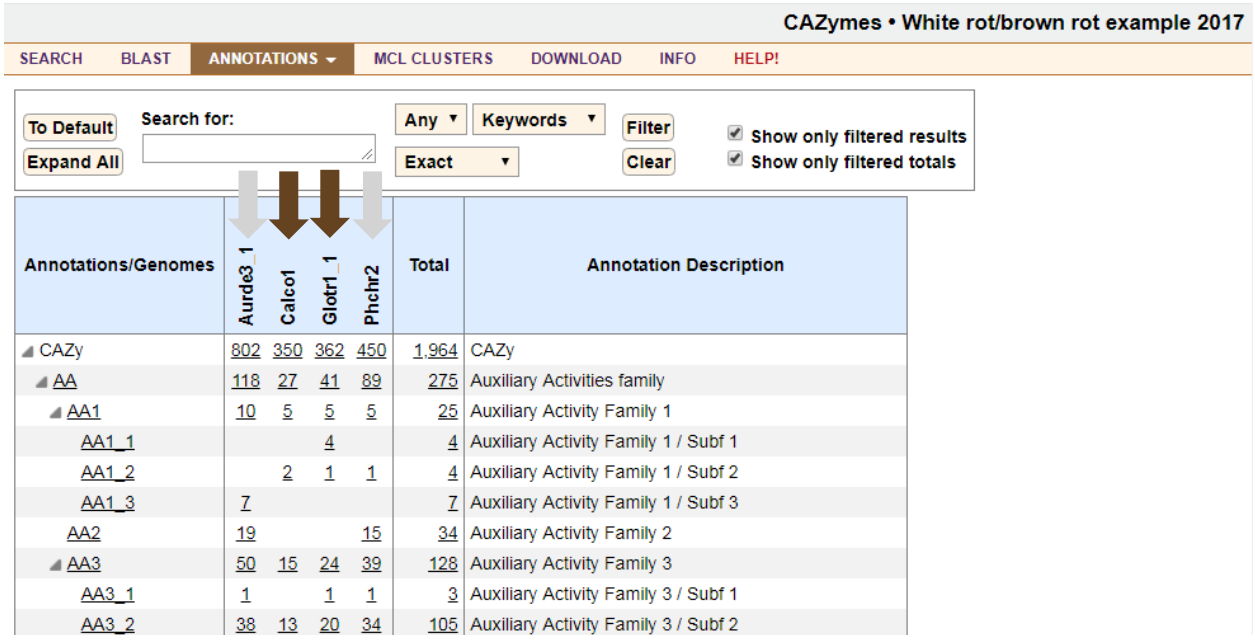


CAZy browser

Click on the CAZYMES item under ANNOTATIONS in the Main menu.



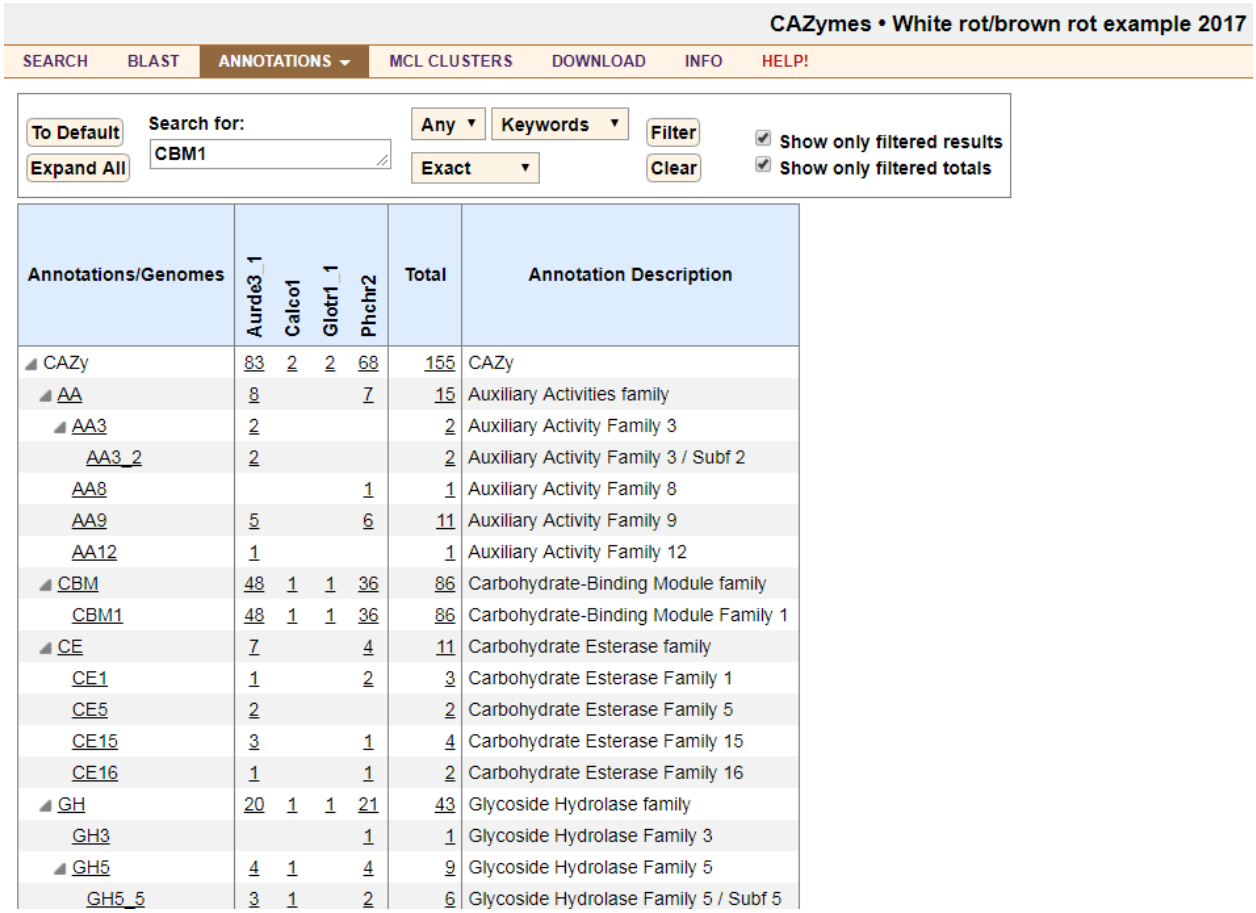
Here you will see a table representation of the predicted CAZymes (Levasseur et al. 2013). The organisms are labeled along the top. The CAZymes are organized by family and labeled along the sides. The numbers in the table tell you how many proteins from each organism's gene catalog were annotated with a given CAZyme. There is also a totals column. Notice that the CAZymes are hierarchically organized: you can see the total number of genes assigned to the general enzyme category (e.g. 'AA'). To expand top level assignment, click on the small arrow left of the catergory, or use the "Expand All" button at the top. Family designations ('AA1', 'AA2', etc.), and to subfamilies ('AA1_1', 'AA1_2', etc.) will then show up. In the image below, arrows have been added in grey for white rot fungi and brown for brown rot fungi.



If we read Levasseur et al. 2013 we know that the AA2 family consists of class II peroxidases that may degrade lignin. Browsing the table, we see that for AA2, we see that *P. chrysosporium* and *A. subglabra* possess 15 and 19 copies of AA2, whereas *G. trabeum* and

C. cornea possess no AA2s. This might suggest that the former two are white rot fungi and the latter two brown rot fungi!

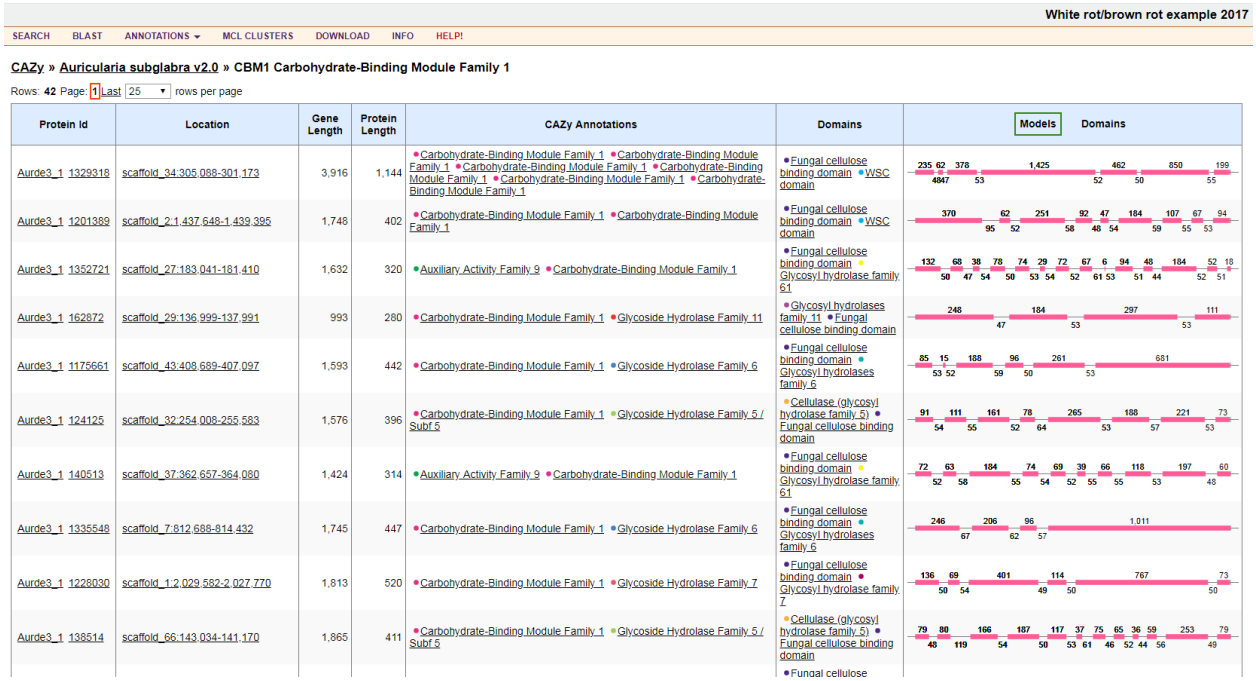
What about the carbohydrate binding motifs, CBM1? Let's say we don't want to scroll through the entire list of CAZymes. Type 'CBM1' into the 'CAZY terms' search box. This will limit the view to only those CAZymes that have a CBM1. Why do so many CAZymes besides CBM1 show up? Because CBM1 co-occurs on the same protein chain with many other CAZymes of diverse function. The numbers in the table will now show, for each CAZyme's row, the number of proteins that also have a CBM1.



Annotations/Genomes	Aurde3_1	Calco1	Glotr1_1	Phchr2	Total	Annotation Description
CAZy	83	2	2	68	155	CAZy
AA	8			7	15	Auxiliary Activities family
AA3	2				2	Auxiliary Activity Family 3
AA3_2	2				2	Auxiliary Activity Family 3 / Subf 2
AA8				1	1	Auxiliary Activity Family 8
AA9	5			6	11	Auxiliary Activity Family 9
AA12	1				1	Auxiliary Activity Family 12
CBM	48	1	1	36	86	Carbohydrate-Binding Module family
CBM1	48	1	1	36	86	Carbohydrate-Binding Module Family 1
CE	7			4	11	Carbohydrate Esterase family
CE1	1			2	3	Carbohydrate Esterase Family 1
CE5	2				2	Carbohydrate Esterase Family 5
CE15	3			1	4	Carbohydrate Esterase Family 15
CE16	1			1	2	Carbohydrate Esterase Family 16
GH	20	1	1	21	43	Glycoside Hydrolase family
GH3				1	1	Glycoside Hydrolase Family 3
GH5	4	1		4	9	Glycoside Hydrolase Family 5
GH5_5	3	1		2	6	Glycoside Hydrolase Family 5 / Subf 5

Notice the abundance of CBM1-encoding genes in *P. chrysosporium* and *A. subglabra*, while *G. trabeum* and *C. cornea* have only a single CBM1-encoding gene each (co-occuring with GH5_5 and GH10 proteins). All of this indicates that we might be looking at two white-rot and two brown-rot fungi.

Click on the number (e.g. 48 for Aurde3_1) to see the the CBM1-containing proteins of A. suglabra in more detail. Notice a variety of CAZymes co-occur with CBM1, including GH5 (various subfamilies), GH6, and many others.

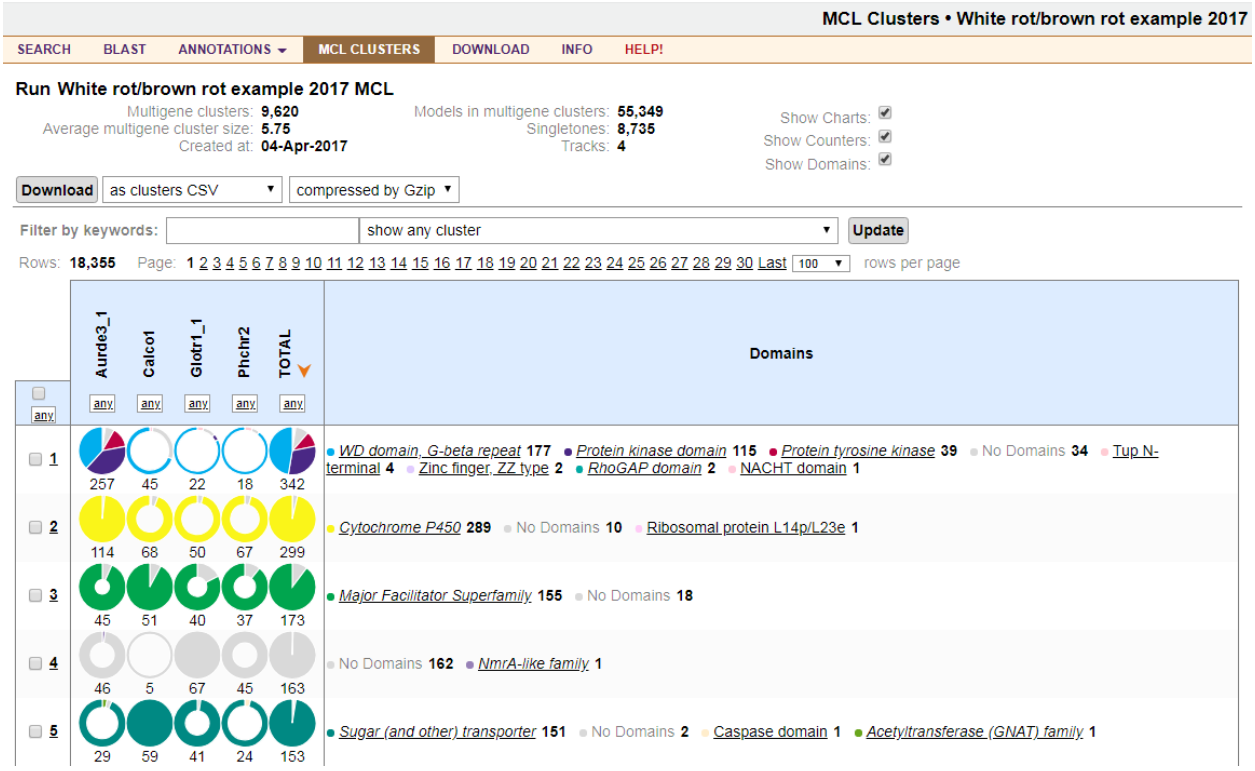


As an exercise, repeat the same search with GH6, GH7, and also the AA9 family of lytic polysaccharide monoxygenases, which may oxidatively act on lignin (Levasseur et al. 2013). Do the presence/absence patterns of these genes indicate the same conclusions about these fungi's mode of decay as we found with AA2 and CBM1? Is it a strict dichotomy, or are there some grey areas in the distribution of these genes?

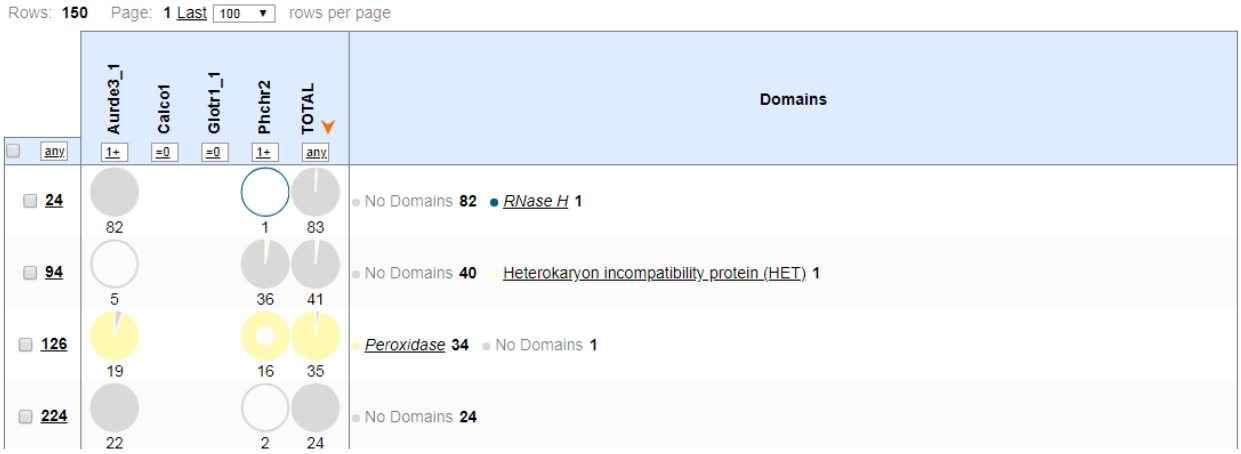
(Answer: *P. chrysosporium* and *A. subglabra* induce white rot wood decay; *G. trabeum* and *C. cornea* brown rot. Notice that brown rot *G. trabeum* has a few AA9 genes, however, indicating that these genes may play a role in brown rot, not just white rot, where AA9s are expanded.)

Cluster page

Now that we have an idea which fungus uses which decay mode, let's ask the reverse question: what are the genes present in one lifestyle, and absent in the other? To do this, click the 'MCL CLUSTERS' item of the Main menu. Here you will see the results of protein sequence clustering by the MCL algorithm (Enright et al. 2002). You can think of clusters as protein families. As with the CAZy browser, the columns indicate organisms. The rows indicate a protein cluster, one cluster per row, with the number of proteins each organism contributes to a cluster. See the HELP Menu for a full explanation of the cluster page.



Notice that under each organism label is a button ‘any’ that can be used to filter clusters by the number of proteins that organism contributes to a cluster, and thus limit which clusters are shown. As an experiment, set the white rot fungi (Aurde3_1 and Phchr2) to “1+” and the brown rot fungi (Calco1 and Glotr1_1) to “=0”. Doing this returns only those clusters which are present in Aurde3_1/Phchr2 and absent in Calco1/Glotr1_1.



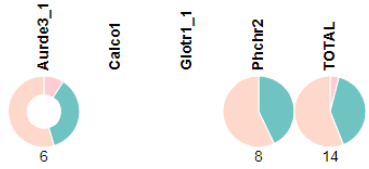
Some 150 clusters fit these criteria. These clusters might include genes important to the white rot decay mode, because they are present in white rot fungi and absent in brown rot fungi. But some of these clusters might have no functional connection to wood decay mode - they are present/absent from the respective kinds of wood decay fungi merely by chance.

These clusters nevertheless represent candidates for further analysis of possible connections to decay mode.

How does one begin interpreting the results? To help with this, each cluster row shows the Pfam domains (http://pfam.xfam.org) that are found in that cluster. Notice that the third row has a "Peroxidase" (PF00141) domain. Notice that the numbers are very close to what we found for the AA2 class II peroxidases in the CAZy browser. It turns out that PF00141 is a superfamily that includes the AA2 enzymes, but it is important to note that not all members of PF00141 can degrade lignin - some have other functions.

Scroll through the rest of the 150 clusters and you will see domains such as Glycosyl hydrolase family 7 and Fungal cellulose binding domain in cluster 507, which roughly overlap with the CAZy GH7 and CBM1 families. Click the '507' to explore that cluster in more detail. On the cluster detail page, a table is presented with one protein per row. Click the 'Domains' view on the rightmost column to see the domain structure of each protein. Notice that all of the proteins have the GH7 domain, and that most, but not all, have a single CBM1 motif at the C-terminus.

Run White rot/brown rot example 2017 MCL » Cluster 507



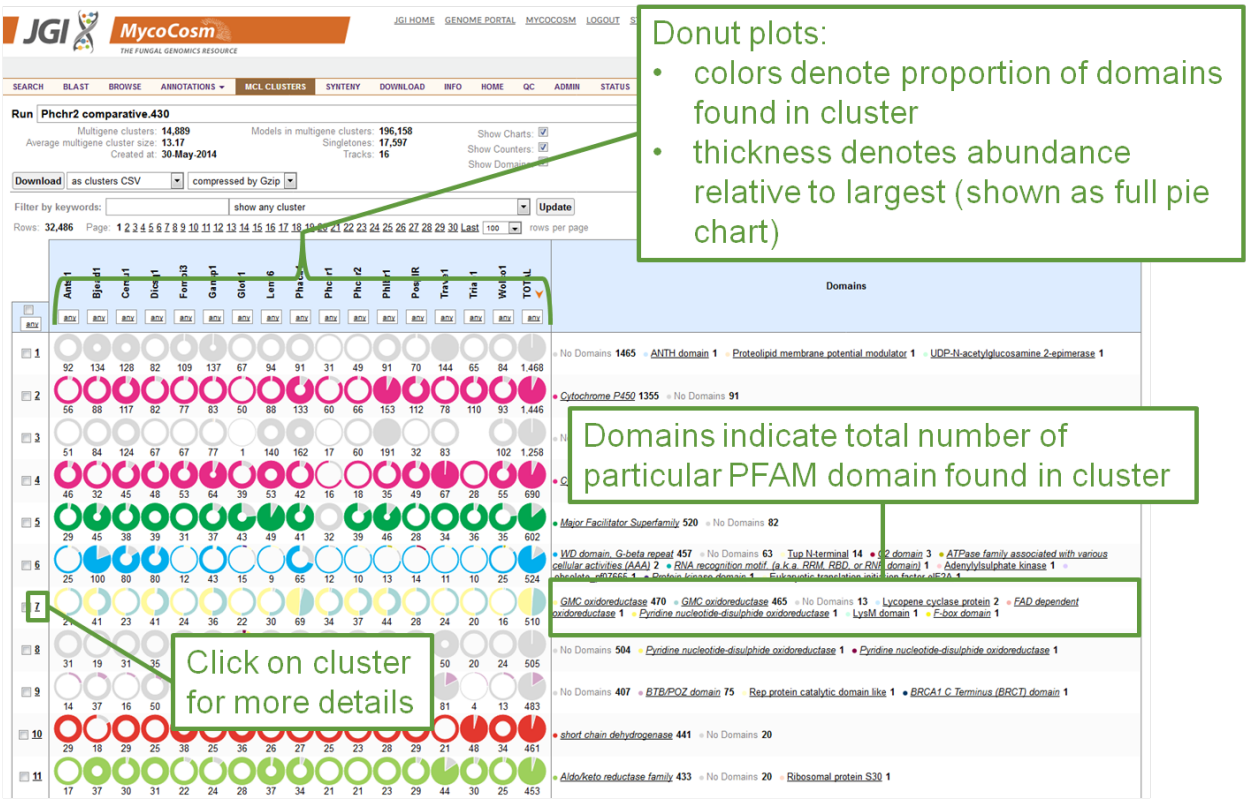
Models: **14** 100 rows per page

Protein Id	Location	Gene Length	Protein Length	Domains	Model / Domains / Synteny
Aurde3_1 1228030	scaffold_1:2,027,770-2,029,582	1,813	520	Glycosyl hydrolase family 7 • *Fungal cellulose binding domain*	136 69 401 114 767 73 / 50 54 49 50 50
Aurde3_1 199579	scaffold_7:522,359-524,267	1,909	515	Glycosyl hydrolase family 7 • *Fungal cellulose binding domain* • Leucine operon leader peptide	139 69 401 114 195 263 294 70 / 52 49 49 57 52 54 51
Aurde3_1 1233301	scaffold_7:538,513-540,388	1,876	519	Glycosyl hydrolase family 7 • *Fungal cellulose binding domain*	125 14 69 401 114 195 566 73 / 229 47 57 52 62 50
Aurde3_1 1310233	scaffold_21:150,095-151,901	1,807	519	Glycosyl hydrolase family 7 • *Fungal cellulose binding domain*	136 69 401 114 764 73 / 48 50 49 50 53
Aurde3_1 1240515	scaffold_21:180,564-182,239	1,676	509	Glycosyl hydrolase family 7	136 470 114 807 / 48 50 51
Aurde3_1 1317126	scaffold_66:202,983-204,748	1,766	449	Glycosyl hydrolase family 7	154 54 137 264 112 137 238 74 177 / 57 49 56 52 52 48 57 48

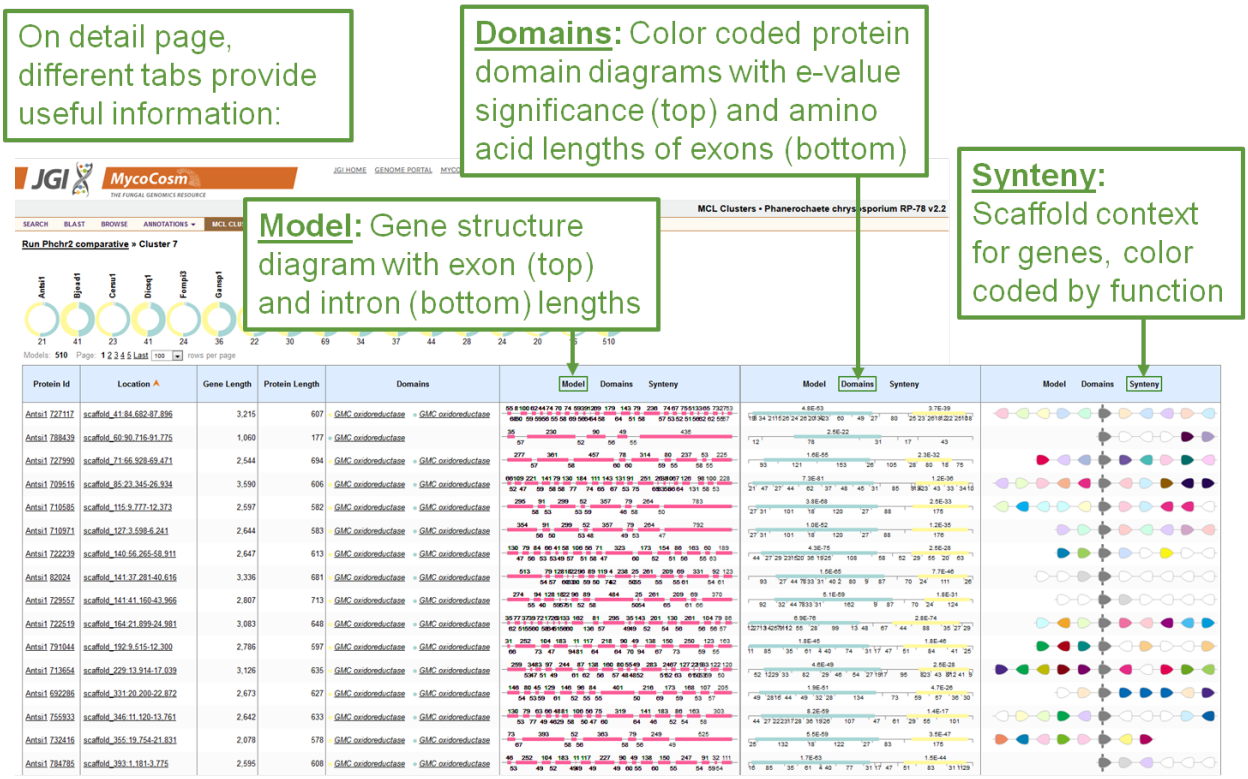
Let's look at what other proteins have the CBM1 carbohydrate-binding motifs in them. Returning to the cluster run page (click the back button, or click the CLUSTERS Menu). Enter the phrase "fungal cellulose binding domain" (be sure to include the quotes) into the "filter by keywords" field. This returns some 26 clusters, all of which have the Pfam domain CBM_1 (PF00734). We see that CBM1 motifs occur in a wide array of domain combinations: often with GMC oxidoreductases, AA9 lytic polysaccharide monoxygenases (formerly GH61), and many hydrolytic enzymes such as GH5, GH6, and GH7. Notice that while these proteins typically are found in expanded copy number in the white rot fungi (Aurde3_1 and Phchr2) they are sometimes found, albeit in lower copy number, in the brown rot fungi (Calco1 and Glotr1_1).

As additional exercises you can (a) search for gene families absent in both white rot fungi; (b) find gene families absent in white rot but present in both brown rot fungi and look at functional domains associated with these families; (c) check if any of these domains are present only in brown rot fungi by resetting filters back to ‘any’ and searching for names of these domains.

A summary of tools available in MCL clustering are shown below.



Clicking in Cluster number provides additional tools as shown below.



References:

- Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R. A., Henrissat, B., Martinez, A. T., Otillar, R., Spatafora, J. W., Yadav, J. S., Aerts, A., Benoit, I., Boyd, A., Carlson, A., Copeland, A., Coutinho, P. M., de Vries, R. P., Ferreira, P., Findley, K., Foster, B., Gaskell, J., Glotzer, D., Gorecki, P., Heitman, J., Hesse, C., Hori, C., Igarashi, K., Jurgens, J. A., Kallen, N., Kersten, P., Kohler, A., Kues, U., Kumar, T. K., Kuo, A., LaButti, K., Larrondo, L. F., Lindquist, E., Ling, A., Lombard, V., Lucas, S., Lundell, T., Martin, R., McLaughlin, D. J., Morgenstern, I., Morin, E., Murat, C., Nagy, L. G., Nolan, M., Ohm, R. A., Patyshakuliyeva, A., Rokas, A., Ruiz-Duenas, F. J., Sabat, G., Salamov, A., Samejima, M., Schmutz, J., Slot, J. C., St John, F., Stenlid, J., Sun, H., Sun, S., Syed, K., Tsang, A., Wiebenga, A., Young, D., Pisabarro, A., Eastwood, D. C., Martin, F., Cullen, D., Grigoriev, I. V., & Hibbett, D. S. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science, 336(6089): 1715-1719.
- Riley, R., Salamov, A. A., Brown, D. W., Nagy, L. G., Floudas, D., Held, B. W., Levasseur, A., Lombard, V., Morin, E., Otillar, R., Lindquist, E. A., Sun, H., LaButti, K. M., Schmutz, J., Jabbour, D., Luo, H., Baker, S. E., Pisabarro, A. G., Walton, J. D., Blanchette, R. A., Henrissat, B., Martin, F., Cullen, D., Hibbett, D. S., & Grigoriev, I. V. 2014. Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. Proc Natl Acad Sci U S A, 111(27): 9923-9928.
- Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M., & Henrissat, B. 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels, 6(1): 41.
- Enright, A. J., Van Dongen, S., & Ouzounis, C. A. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, 30(7): 1575-1584.