Evolutionary and Comparative Fungal CAZomics.

A-5

Annotation, comparative and evolutionary analysis of fungal Carbohydrate Active enZymes (CAZymes).

Etienne G.J. Danchin, Corinne Rancurel, Pedro M. Coutinho, Bernard Henrissat.

etienne.danchin@afmb.univ-mrs.fr [http://edanchin.free.fr]

FUNGI

The most abundant source of available whole genomes in Eukaryotes (about 40 genomes already available, > 50 more in the short term).

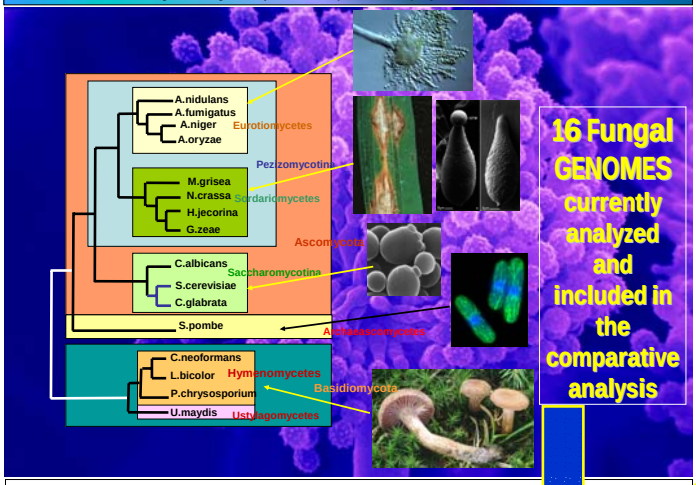
A huge Biodiversity (~ 75,000 species already described, upper estimate of 1.5 million species [1,2].

High variability and diversity of lifestyles (sexual / asexual…), lifeforms (uni- pluri cellular) and ecotypes (pathogens, parasites, saprophytes, symbionts…).

Available Fungal genomes already cover a diverse sample of the huge Biodiversity.

Fungi have close relationships with carbohydrates and are known to be excellent source of CAZymes in particular for industrial applications.

[1] Hawksworth DL: **The fungal dimension of biodiversity: magnitude, significance, and conservation.** *Mycological Research* 1991, **95**:641-655.
[2] Hawksworth DL: **The magnitude of fungal diversity: the 1.5 million species revisited.** *Mycological Research* 2001, **105**:1422-1432.



16 Fungal GENOMES currently analyzed and included in the comparative analysis

CAZymes (Carbohydrate Active enZymes)

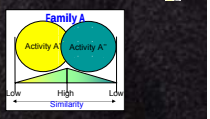
- Enzymes forming, modifying and degrading glycosidic bonds.
- Found in all kingdoms of life (including viruses)
- Particularly important for the biology of species having close relations with sugars (i.e. Fungi).
- Widely used in the industry (food, agriculture, pulp & paper, biomass conversion, etc)
- > 30 % of world enzyme's market.
- Involved in human and animal pathologies.
- Typically 2-3 % of a genome in Eukaryotes (with considerable variations).

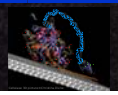
A given CAZyme family Can Feature :

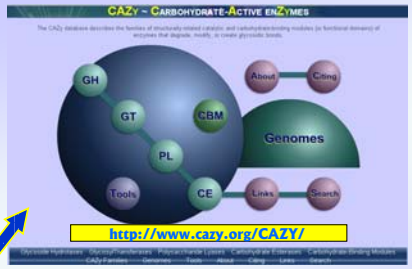
Variability of modular structure

Variability of subtract / product specificity

→ Classical structural and functional annotation based on best BLAST hits is not efficient in that case and is often misleading.







CAZy DB

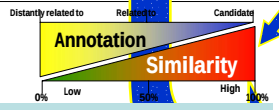
>50,000 curated entries corresponding to candidate CAZymes.
Classified into families :

- **106** Glycoside-Hydrolases (**GH**).
- **87** Glycosyl-Transferases (**GT**).
- **18** Polysaccharide-Lyases (**PL**).
- **14** Carbohydrate-Esterases (**CE**).
- **45** ancillary Carbohydrate Binding Modules (**CBM**).

CAZy Annotation

Structural Annotation takes into account the modular organization. Assignment to families of CAZymes as a function of the enzyme's modular composition and significance of similarity. Based on the expert curated CAZy database containing a library of > 100,000 annotated modules.

Functional annotation designed to "age well". Designed to survive experimental characterization, our functional annotations avoid over-predicted subtract / product specificity. It is based on a gradient as a function of similarity to characterized hits. Only 100% identity to a traceable experimentally characterized hit allows assignment of an E.C. number in CAZy.



Example:
in a family that contains β-mannosidases, β-galactosidases and β-glucuronidases, all enzymes hydrolyze equatorially oriented glycosidic bonds. A strong similarity to β-galactosidases allows annotation as "candidate β-galactosidase", but if similarity is not sufficient for a safe prediction of substrate specificity, the best possible annotation is "candidate β-glycosidase". If the similarity is even weaker the best possible annotation would be "related to β-glycosidase".

FUNGAL CAZomes

> 155,000 **ORFs** scanned against CAZy libraries.
> 4,700 genes encoding putative **CAZymes**, including:

~ 2,600	**GHs**	covering	63 families
~ 1,300	**GTs**	covering	35 families
~ 450	**CBMs**	covering	15 families
~ 310	**CEs**	covering	9 families
~ 105	**PLs**	covering	8 families

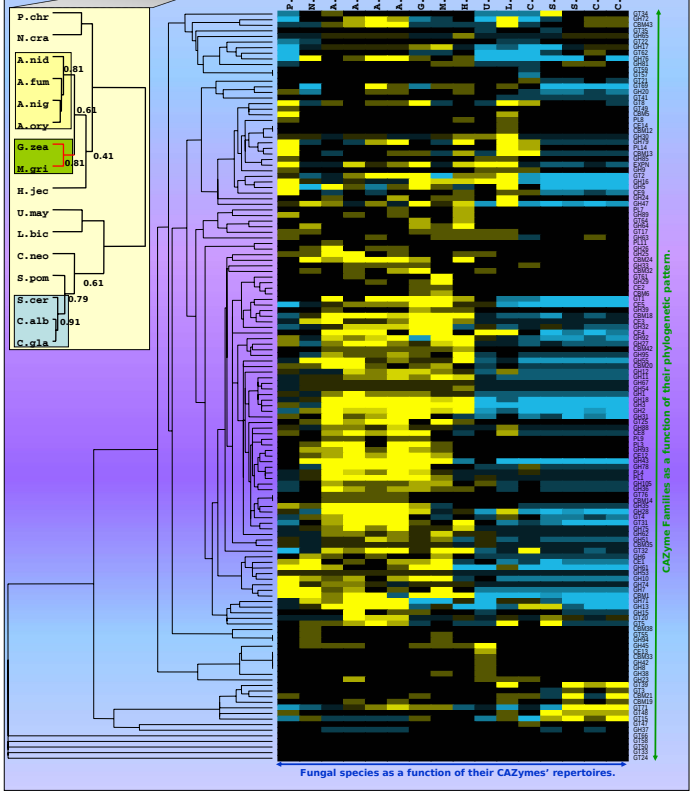
On average, **3% of a Fungal Genome** (among the **higher rates** in eukaryotes)

Wide variations from ~2 % in Saccharomycotina to ~ 4 % in Aspergilli.

 # GHs > # GTs in Fungi as **opposed** to other Eukaryotes.

Wider variations of the **GH** repertoires (SD ~4.25) than the GT's (SD ~2.93). **More** universally **conserved GT** families than **GH** families.

Double hierarchical clustering:



CAZyme repertoires vary both as a function of the **phylogenetic distance** and as a function of **lifestyle / ecotype**.

The **GH families**, a significant proportion of which are **secreted**, show **higher variations** and seem to be more subject to **adaptations** to the environment. Polysaccharide Lyases (**PL**) follow the **same trend**.

In contrast, the **GT** repertoires appear more "**housekeeping-like**" and their divergence reflects more the phylogenetic distance than the lifestyle.

Consistent with previous observations suggesting that **protein secretion is coupled to higher evolutionary rates** [1] (at amino-acid substitution level).

[1] Luz H, Vingron M: **Family specific rates of protein evolution**. *Bioinformatics* 2006, **22**(10):1166-1171.

	Proteome size	GH	%GH	GT	%GT	CBM	PL	CE	% CAZymes
A.nid	9 541	247	2.59	89	0.93	34	19	28	4.01
A.fum	9 926	260	2.62	95	0.96	51	13	28	3.99
A.nig	14 165	240	1.69	111	0.78	35	8	23	2.70
A.ory	12 074	283	2.34	110	0.91	29	21	27	3.65
M.gris	11 109	229	2.06	87	0.78	56	4	47	3.30
N.cra	10 079	169	1.68	71	0.7	38	3	21	2.62
H.jec	9 997	193	1.93	87	0.87	35	3	15	2.98
G.zea	11 640	237	2.04	95	0.82	60	20	42	3.38
C.alb	6 419	58	0.9	69	1.07	4	0	3	2.03
S.cer	6 294	45	0.71	67	1.06	9	0	3	1.83
C.gla	5 283	38	0.72	73	1.38	10	0	3	2.16
S.pom	4 824	48	1	61	1.26	5	0	5	2.36
C.neo	6 572	75	1.14	62	0.94	9	3	9	2.27
L.bico	21 244	163	0.77	83	0.39	24	7	19	1.28
P.chr	11 777	179	1.38	65	0.7	44	4	16	2.24
U.may	6 522	101	2.74	71	1	8	1	20	2.96

Ascomycetes

Richest sets of **degrading** CAZymes (GHs, PLs, CEs) in **Versatile Saprophytes** and **Phytopathogens**

Relatively "**constant**" sets of **GTs**

Basidiomycetes

~ **Rich** sets of **degrading** CAZymes (GHs,CEs) in a **Saprophyte**, a **Symbiont**, and a Phytopathogen (CEs only)

"**Constant**" set of **GTs**

Variable set of **CBMs**

Paucity of **PLs**

Work in Progress: Identification of CAZyme families that underwent significant **expansions / reduction during evolution**. Collaboration with Matthew Hahn, University of Indiana [1,2]. Early results confirm that **GH families evolve faster** than GT families. We identified several CAZyme families in each class that significantly expanded or reduced in specific lineages.

[1] Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N: Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res 2005, 15(8):1153-1160.
[2] De Bie T, Cristianini N, Demuth JP, Hahn MW: CAFE: a computational tool for the study of gene family evolution. Bioinformatics 2006, 22(10):1269-1271.

Future directions:

- **Classify species** based on **subset** of CAZymes families and subfamilies known to contain members **involved** in activities of **industrial** or **biological interest** (pectinolytic, cellulolytic, …)
- Correlate **clustering** of fungal **species** based upon **CAZyme repertoires** to actual **experimental** biological **data** related to their ability to degrade various substracts and other biological features.

Glycogenomics and Biomedical Structural Biology, AFMB UMR 6098, CNRS, Université de Provence, Université de la Méditerranée, 163 Av. de Luminy, Case 932, 13288 Marseille Cedex 09. France.

http://www.afmb.univ-mrs.fr/

STREP fungwall European Consortium